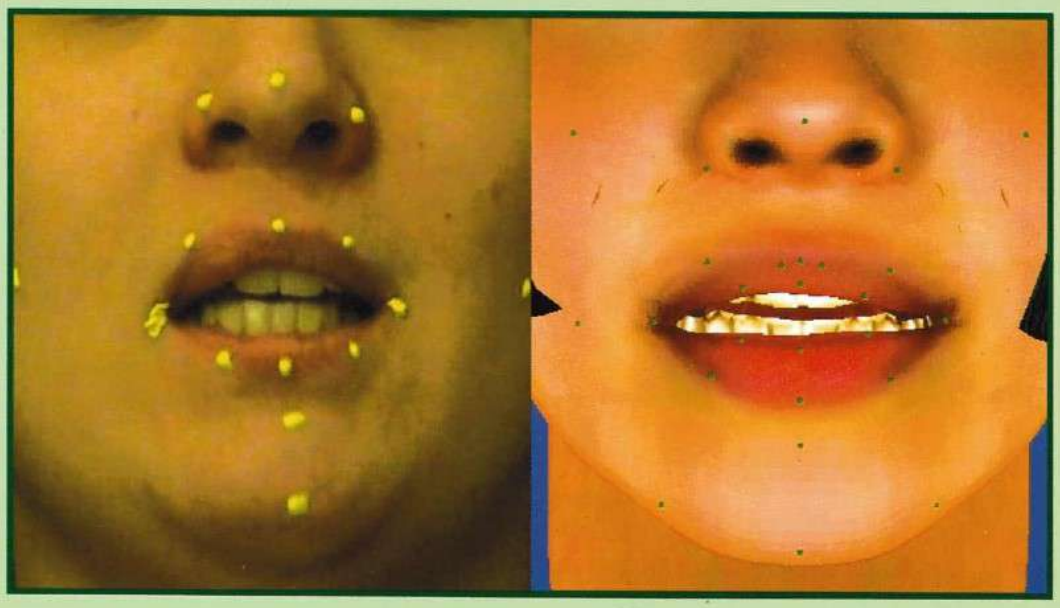
híradástechnika

1945 VOLUME LXI. 2006

hírközlés ▪ informatika



Szolgáltatás-automatizálás és beszédtechnológia

Jelölőnyelvek

Szolgáltatások

Beszédtechnológiák

2006/3

A Hírközlési és Informatikai Tudományos Egyesület folyóirata
a Nemzeti Hírközlési és Informatikai Tanács együttműködésével

Tartalom

Címlap: Vizuális beszédszintézis adatgyűjtéséhez megjelölt referenciapontok természetes és mesterséges arcmodellen (Fotó: Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar)

Szolgáltatás-automatizálás és beszédtechnológia

nemeth@tmit.bme.hu
szabo@hit.bme.hu

Ebben a számunkban azt kívánjuk bemutatni, hogy a beszédtechnológia eredményei hogyan jelennek meg az infokommunikációs szolgáltatások palettájának bővítésében és azok minőségének javításában.

Az első oldalakon egy áttekintő jellegű cikket olvashatnak a beszédtechnológiai alkalmazások gyors és hatékony fejlesztését támogató jelölőnyelvekről.

Az ezt követő blokkban a végfelhasználók számára remélhetőleg rövid időn belül elérhető szolgáltatásokat tárgyaló cikkek találhatók.

A gyógyszerek felhasználói utasítását teljesen automatizáltan – specializált beszédfelismerő és szövegfelolvasó felhasználásával – elérhetővé tevő rendszer ismertetése jó példa arra, hogy a mégoly triviálisnak tűnő emberi funkciók sikeres gépi megoldásához is milyen sokrétű elemző és alkotó munka szükséges.

Egy másik új fejlesztés az orvosok munkájának hatékonyságát javíthatja. A bemondás alapján az írott leletet automatikusan elkészítő rendszer bemutatása jól jelzi a formalizált és a természetes kommunikáció eltéréséből adódó nehézségeket.

A szövegfelolvasó rendszerek fejlődését szemlélteti az időjárás-jelentések témakörén keresztül egy kiemelkedően alapos teszteléssel alátámasztott cikk.

Szellemes és újszerű ötlet az akusztikus jelből a szájmozgás vizuális paramétereit közvetlenül meghatározó eljárás, ami nagy segítséget jelenthet egészséges és siket emberek közvetlen kapcsolatteremtésében.

A harmadik részben a különböző beszédtechnológiai alkalmazások kifejlesztéséhez szükséges technológiai elemekről és háttérmegoldásokról olvashatunk.

Bemutatásra kerülnek azok a nyelvstatisztikai és elemzési szempontok és módszerek, melyek felhasználásával olyan szövegtörzs alakítható ki, melynek felolvasása és feldolgozása után az adott témakörre az emberi minőséget megközelítő gépi felolvasás állítható elő.

Az interdiszciplináris megközelítés szükségességét és jelentőségét jól illusztrálja a korpusz alapú beszédszintézis nyelvészeti és fonetikai kérdéseit ismertető írás.

A szövegfelolvasó rendszerek a szöveg előfeldolgozását valós időben végző modulok nélkül gyakorlati alkalmazásokba nehezen helyezhetők. Ilyen probléma például a szöveg nyelvének megállapítása vagy a szófajok meghatározása. Egy erre a témakörre kifejlesztett gépi tanuló algoritmust is ismertetünk.

Beszédtechnológiai számunkat egy klasszikus téma – a beszéd detektálása, egyéb jelektől és szünettől való megkülönböztetése – újszerű, a témakör ETSI szabványánál jobb eredményeket elérő megközelítéséről szóló beszámoló zárja.

Németh Géza,
vendégszerkesztő

Szabó Csaba Attila,
főszerkesztő

Hangos jelölőnyelvek
Jelölőnyelvek a beszéd alapú alkalmazások fejlesztésében

ABARI KÁLMÁN

Debreceni Egyetem, Pszichológiai Intézet és Matematikai-Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu

Lektorált

Kulcsszavak: beszéd alapú alkalmazás, szabványok, SALT, SRGS, SSML, VoiceXML

Az utóbbi években a beszéd alapú alkalmazások fejlesztésében az egyéni megközelítések helyét fokozatosan az ipari szabványokon alapuló stratégiák és architektúrák veszik át. Különösen igaz ez a telefonos és a multimodális alkalmazásokra, melyek fejlesztését mára majd egy tucat XML alapú jelölőnyelv segíti. A cikkben összefoglaljuk a beszéd alapú alkalmazások egyes komponenseit és azok kommunikációját leíró jelölőnyelveket.

1. Bevezetés

Az elmúlt évek hatalmas technológiai fejlődése ellenére a beszéd alapú alkalmazások fejlesztése összetett feladat, hiszen olyan bonyolult technológiák integrációjára van szükség, mint például a beszédfelismerés, beszédszintézis és dialógusvezérlés. A régebbi alkalmazások elsődlegesen fejlesztőik egyéni megoldásain alapultak, habár a különböző nyílt programozási felületek (API-k) megjelenése – például SAPI (Microsoft Speech Application Program Interface), JSAPI (Java Speech API) – jelentősen csökkentette az alkalmazásfejlesztés bonyolultságát.

Az 1990-es évek végétől aztán egy igen kedvező folyamat indult el: az egyéni megközelítések helyét fokozatosan az ipari szabványokon alapuló stratégiák és architektúrák veszik át. Ennek a szabványosítási folyamatnak a legjelentősebb hajtómotorja a webes és a telefonos világ összekapcsolásának igénye volt. Az áhított cél, hogy ugyanazok a szolgáltatások, amelyeket az ügyfelek eddig hagyományosan grafikus felületről értek el, ezután telefonon keresztül, a meglévő webes infrastruktúrával együttműködve, hang alapú kérések formájában is hozzáférhetőek legyenek. Az integrációs törekvés szimmetrikus, tehát az a cél, hogy az adatbevitel grafikus és hangalapú módon egyaránt megtörténhessen. Ennek érdekében az utóbbi nyolc évben majd egy tucat jelölőnyelvet fejlesztettek ki, melyek a beszéd alapú alkalmazások egyes részeinek szabványos leírását teszik lehetővé. E cikkben ezeket a „hangos” jelölőnyelveket tekintjük át.

2. Testületek

A szabványok alkalmazása a beszéd alapú alkalmazások fejlesztésében – azon túl, hogy jelzik, a terület kezd nagykorúvá válni – számos előnnyel jár. Elrejtik a technológiai részleteket, biztosítják a különböző szállítóktól érkező komponensek együttműködését, kevesebb időbefektetés és kisebb erőfeszítés mellett újrafelhasználható és hordozható megoldások létrehozását támogatják. Másfelől azonban a fejlesztők korlátozva érezhetik a kreativitásukat és bosszankodhatnak, ha valamely funkciót az adott szabvány (még) nem támogatja.

Szabvány alatt a továbbiakban olyan leírást értünk, melyet valamely szabványosításért felelős testület formálisan elismert. A beszéddel kapcsolatos területen a következő szervezetek a legaktívabb:

• A **W3C (World Wide Web Consortium)** hagyományosan vezető szerepet játszik a webes technológiák kifejlesztésében, a Webben rejlő lehetőségek minél teljesebb kihasználásában. Az egyes specifikációk kidolgozása munkacsoportokban történik, melyet a W3C tagjai alkotnak. Egy többlépcsős folyamat eredménye (munkaterv, utolsó felhívás munkatervre, előzetes javaslatterv, javaslatterv, ajánlás) míg egy specifikációból W3C-ajánlás lesz, amelyre a webes társadalom és az ipar már szabványként tekint [5]. A beszéd és multimodális alkalmazások területén két munkacsoport végez fejlesztést, a Voice Browser Working Group (Hangböngésző Munkacsoport) és a Multimodal Interaction WorkGroup (Multimodális Interakció Munkacsoport).

• Az **IETF (Internet Engineering Task Force)** célja az Internet működésének és fejlődésének előmozdítása, az egyes protokollok használatának szabályozása. A Speech Services Control (SpeechSC) munkacsoport az elosztott környezetben működő biztonságos beszédfeldolgozás szabványaiért felelős.

• A **ETSI (European Telephony Standards Institute)** célja azon szabványok kidolgozása, amelyek biztosítják, hogy a globális távközlési piac egyetlen piacként működjön. Az Aurora projekt a mobilhálózaton megvalósuló elosztott beszédfelismerés szabványosításán dolgozik.

Két további vállalati összefogáson alapuló „fórum” is meghatározó szerepet játszik ezen a területen:

• A **VoiceXML Forum** olyan nagyvállalatok összefogásából alakult ki, melyek mindegyikének korábban megvolt a saját ötlete a hang alapú webes szolgáltatásra. Ez az AT&T és a Lucent Technologies vállalatok PML specifikációja, a Motorola SpeechML-je és az IBM Vox-

ML-je volt. Mivel érdekeltek voltak az egységes hangvezérelt Web létrehozásában, közösen elkészítették a VoiceXML 1.0-s változatát, amit 2000 márciusában bemutattak a W3C-nak [9]. Azóta a fórum nem vesz részt a nyelv továbbfejlesztésében, munkája az oktatásra és a webes technológiák népszerűsítésére korlátozódik.

• A **SALT Forum**, amely a Cisco, Comverse, Intel, Microsoft, Philips és Scansoft összefogásából jött létre 2001-ben. Közösen dolgozták ki a SALT (Speech Application Language Tags) 1.0-s változatát, melyet 2002-ben bemutattak a W3C-nak [8].

3. Architektúrák

A Web által kínált információk hagyományos elérési módja a személyi számítógépek grafikus felülete, mely a kommunkáció során a „rámutatás" (point and click) elvet követi, néha a billentyűzetet használja adatbevitelre. A hang alapú interfész ehhez képest a mindennapos beszélgetésben megszokott, „beszélek és hallgatok" stílust követi, olyan eszközöket felhasználva, mint szóbeli utasítok, előre felvett beszéd visszajátszása, szintetizált beszéd, és szükség esetén a telefonok nyomógombjai. Irodai környezetben a vizuális felület használata a leghatékonyabb, ahol rendelkezésre áll szélessávú átviteli csatorna, nagyfelbontású képernyő, egér és billentyűzet. A hang alapú felület akkor a leghasznosabb, amikor távol vagyunk az íróasztalunktól, illetve egyes speciális felhasználói csoportoknak, mint például a látássérültek és látáskorlátozottak. Ha tehát a webes szolgáltatások univerzális elérését akarjuk biztosítani, akkor mindkét megközelítési módot, a vizuális és a hang alapú felületet is támogatnunk kell.

Négy alapvető módszert ismerünk, melyek segítségével grafikus és hang alapú felhasználói felület is biztosítható webes alkalmazásunkhoz:

– **Különállóan megtervezett grafikus- és hanginterfész,** melyek ugyanazokra az adatokra és üzleti logikára támaszkodnak, de egymástól függetlenül lettek kifejlesztve.
– **A hagyományos vizuális böngésző „meghangosítása",** mely során grafikus böngészőnk az épp megjelenített lap tartalmát fel tudja olvasni, és szóbeli utasítások segítségével oldalak közötti navigációra is képes.
– **Átkódolás (transcodig),** mely során a meglévő (X)HTML dokumentumokból automatikusan állítunk elő hang alapú interfészt.
– **Kombinált grafikus- és hanginterfész,** ahol minden egyes oldal tartalmaz a grafikus és a hang alapú felhasználói felületre is információt. Ez nem multimodális interfészt jelent, hiszen egyszerre csak az egyik modalitás használható.

A kombinált grafikus- és hanginterfészt bonyolult tervezés, implementálás és karbantartás jellemzi, a „meghangosított" vizuális böngésző és az átkódolási technika esetén pedig nehezen biztosítható a vegyes kezdeményezésű dialógusvezérlés. Ezeket a hátrányokat a grafikus felülettől függetlenül megtervezett hanginterfészek kiküszöbölik, így nem érzékenyek a vizuális interfész változására és a vezérlés jellege is tetszőlegesen megválasztható. A következő pontban ismertetendő szabványok és nyílt specifikációk az ilyen különállóan megtervezett hangalapú felületek fejlesztését támogatják, melyeket többnyire a nyomógombos bevitelt beszédfelismeréssel kombináló *telefonos alkalmazások* körében használhatjuk. Néhány szabvány *multimodális alkalmazások* létrehozását is támogatja, melyek a beszéd feldolgozásán túl az olyan hagyományos perifériák párhuzamos használatát biztosítják, mint például egér, billentyűzet és képernyő.

Rövidítések

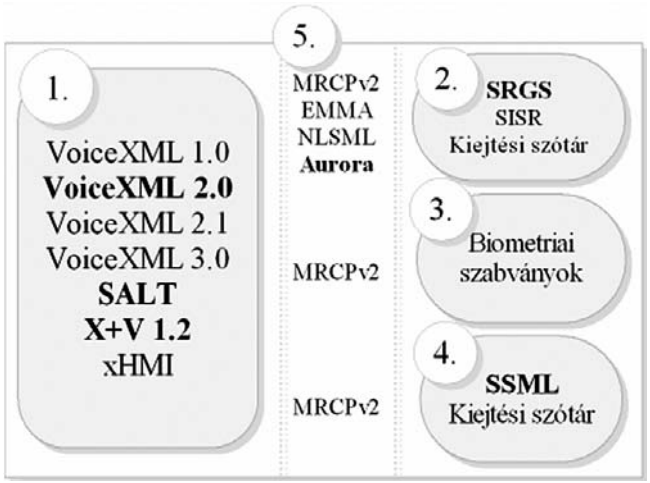
API	Application Programming Interface
ECMA	European Computer Manufacturers Association
EMMA	Extension Multi-Modal Annotation
ETSI	European Telephony Standards Institute
DSR	Distributed Speech Recognition
HTML	Hypertext Markup Language
IETF	Internet Engineering Task Force
JSAPI	Java Speech API
JSGF	Java Speech Grammar Format
JSML	Java Speech Markup Language
MRCP	Media Resource Control Protocol
NLSML	Natural Language Semantics Markup Language
SALT	Speech Application Language Tags
PML	Phone Markup Language
SAPI	Microsoft Speech Application Program Interface
SISR	Semantic Interpretation for Speech Recognition
SMIL	Synchronized Multimedia Integration Language
SRGS	Speech Recognition Grammar Specification
SSML	Speech Synthesis Markup Language
SVG	Scalable Vector Graphics
TTS	Text-to-speech
W3C	World Wide Web Consortium
VoiceXML	Voice Extensible Markup Language
X+V	XHTML+Voice
xHMI	Extensible Human-Machine Interface
XHTML	Extensible Hypertext Markup Language
XML	Extensible Markup Language

4. Szabványok és nyílt specifikációk

A beszéd alapú alkalmazások fejlesztését lehetővé tevő különböző szabványokat és nyílt specifikációkat két csoportba sorolhatjuk: az alkalmazás *leírására* hasz-

nálatos szabványok (1. ábra: 1-4) illetve az így elkészült szoftver komponensek közötti *kommunikációt* elősegítő specifikációk (5). Az alkalmazás leírásához használt nyelvek további csoportjai – tükrözve a beszéd alapú alkalmazás általános felépítését – a dialógusvezérlés (1), a bemenő és kimenő beszéd kezelése (2 és 4), valamint a beszélő azonosítás (3) funkciókat fedik le. Az *1. ábrán* kiemelve a ma használatos, teljesen kidolgozott szabványok vagy nyílt specifikációk szerepelnek, a többi fejlesztés alatt áll, kivéve a VoiceXML 1.0 és az NLSML, melyek túlhaladott szabványok.

1. ábra Beszédtechnológiai specifikációk
1.– dialógusvezérlés; 2.– beszéd bemenet;
3.– beszélő azonosítás;
4.– beszéd kimenet; 5.– kommunikáció



4.1. Dialógusvezérlés

A dialógusvezérlés felelős a teljes beszédfolyamat vezérléséért, a felhasználóval való kommunikációért. A dialógusvezérlés dönti el, hogy a rendszer mikor mit mondjon, illetve mikor figyelje a felhasználó szóbeli utasításait, és milyen válaszokra számítson. Ő ad utasításokat a bemenő és kimenő beszédért valamint a beszélő azonosításáért felelős komponenseknek.

A dialógusvezérlőknek többféle megközelítése létezik, de a napjainkban használt szabványosnak tekinthető megoldások a webes paradigmát követik. Azaz a webszerver jóldefiniált jelölőnyelven írt lapokat küld a böngészőnek, ha az kéri, amiket aztán a böngésző értelmez és végrehajt. A legfontosabb dialógusvezérlő jelölőnyelvek: VoiceXML, SALT, X+V, xHMI.

VoiceXML

A legrégebbi és a legtöbbet hivatkozott szabványos dialógus leíró formanyelv a VoiceXML (rövidebben VXML), aminek az 1.0-ás változatát még 2000-ben a VoiceXML Forum definiálta. Ebből a változatból indult a W3C Hangböngésző Munkacsoportja és készítette el a mára ajánlássá vált VoiceXML 2.0-t (2004. március). A cikk írásának idején a VoiceXML 2.1 „felhívás utolsó munkatervre" fázisban van, a 3.0-ás változatnak pedig az előkészítése folyik.

Egy VoiceXML alkalmazás általában több dokumentum együttese, ezek Web-szerveren tárolódnak, vagy szerver oldali szkriptek generálják őket. A VoiceXML böngésző dokumentumokat tölt le, értelmezi őket, majd inputot kér a felhasználótól és figyeli a választ. Azt az időtartamot, míg a felhasználó kapcsolatban van a VoiceXML böngészővel, ügymenetnek (session) nevezzük. Egy ügymenet során a hangböngésző általában több VoiceXML dokumentumot futtat. Egyidőben két VoiceXML dokumentum lehet aktív, az egyik a gyökér dokumentum (root document), mely az alkalmazásban mindig aktív, a másik a gyermekdokumentum, ami az alkalmazás egy részletét tartalmazza. Az aktív gyermekdokumentum az alkalmazás működése során mindig cserélődik.

Két elsődleges vezérlő van a VoiceXML-ben: a menü (menu) és az űrlap (form). A menü általában egy prompt lejátszást és a felhasználó szóbeli utasításának figyelését jelenti. Amikor felhasználó normál beszéd segítségével kiválasztja, hogy merre akar továbbmenni az alkalmazásban, akkor arról dönt, hogy melyik dokumentum töltődjön le és vegye át a gyermekdokumentum szerepét. Az űrlap mezőket (field) tartalmaz, melyek szóbeli közléseink alapján értéket kapnak. A mezők „kitöltését" hangos üzenetek (block) megszólaltatásával segíthetjük, és mezők kitöltöttségét is tudjuk ellenőrizni (filled). Az Űrlap Értelmező Algoritmus (Form Interpretation Algorithm, FIA) felelős a soron következő mező kiválasztásáért, a mezők kitöltését pedig nyelvtanok (grammar) felügyelik. A kitöltési algoritmus normális működését események (event) és az azokat lekezelő programrészek (event handler) futásai szakítják meg időlegesen.

Az *1. példa* egy prompt lejátszással kezdődik (4-6. sor), majd a felhasználó szóbeli választásának megfelelően (7-12. sor), az adott űrlapra lépve (14-16. vagy 17-19. sor), az alkalmazás prompt lejátszással (15. vagy 18. sor) nyugtázza döntésünket:

1. Példa
Egy VoiceXML menü

```
1  <?xml version = "1.0"?>
2  <vxml version = "2.0">
3  <menu id="travel">
4  <prompt>
5    Do you want to travel by rail, or boat?
6  </prompt>
7  <choice next="#train">
8    rail
9  </choice>
10 <choice next="#boat">
11   boat
12 </choice>
13 </menu>
14 <form id="train">
15  <block> You have selected to travel by rail.</block>
16 </form>
17 <form id="boat">
18  <block> You have selected to travel by boat.</block>
19 </form>
20 </vxml>
```

A *2. példában* az induló prompt lejátszás (5-8. sor) az űrlapon szereplő egyetlen mező (4-19. sor) kitöltésére szólít fel, amit az adott nyelvtannak (9-18. sor) megfelelően (értéke csak „march", „april", vagy „may" lehet) kell elvégeznünk. A sikeres kitöltés nyugtázását (20-22. sor) a *monthofyear* változó használata jelentősen leegyszerűsíti.

```
1  <?xml version = "1.0"?>
2  <vxml version = "2.0">
3  <form id="checkmonth">
4   <field name="monthofyear">
5    <prompt>
6     Please say the name of any month
7     from march to may.
8    </prompt>
9    <grammar type="application/srgs+xml"
10      root="monthofyear">
11    <rule id="monthofyear" scope="public">
12     <one-of>
13      <item>march <tag>march</tag></item>"
14      <item>april <tag>april</tag></item>"
15      <item>may <tag>may</tag></item>"
16     </one-of>
17    </rule>
18   </grammar>
19  </field>
20   <block>
21    You have chosen <values expr="monthofyear" />
22   </block>
23  </form>
24  </vxml>
```

2. Példa
Egy VoiceXML űrlap

A VoiceXML támogatja továbbá aldialógusok (subdialog) használatát gyakran ismétlődő részek kényelmes felhasználására, változók létrehozását, melyekkel például az aldialógusokat paraméterezhetjük, és az ECMAScript-et, mellyel procedurális feldolgozást végezhetünk.

SALT

A Speech Application Language Tags (SALT), amit a SALT Forum 2001-ben tett közzé, multimodális és telefonos alkalmazások fejlesztését is támogatja.

A SALT nyílt specifikáció néhány XML jelölő együttese, melyeket olyan gazdanyelvekbe ágyazhatunk, mint az XHTML, SVG, SMIL.

A legfontosabb jelölők a következők:

<prompt> előre felvett vagy szintetizált beszéd lejátszásáért felelős,
<listen> a felhasználó szóbeli utasításait figyeli,
<grammar> a felhasználó lehetséges közléseiben szereplő szavakat, kifejezéseket írja le,
<dtmf> a telefonos alkalmazások számára nyomógombos bevitelt ír elő,
<record> hangfelvételt tesz lehetővé,
<bind> a felhasználótól származó, felismert közléseket integrálja az üzleti logikával.

A SALT nem rendelkezik vezérlésátadó funkciókkal, azokról a gazdanyelvnek kell gondoskodnia. A *3. példa* egy üdvözlő prompt lejátszással kezdődik (6-9. sor), majd ha az befejeződött (*oncomplete* jellemző), újabb prompt lejátszás (10-12. sor) és a felhasználó figyelése (13-16. sor) következik. A jelszó megadása után a *<bind>* elem hatására az *iptPIN* beviteli mező kitöltésre kerül (15. sor).

```
1  <html xmlns:salt="http://www.saltforum.org/02/SALT">
2    <body onload="sayWelcome.Start()">
3     <form id="PIN" action="checkPIN.html">
4           <input id="iptPIN" type="text" />
5     </form>
6     <salt:prompt id="sayWelcome" oncomplete=
7     "askPIN.Start(); recoPIN.Start()">
8        Welcome to my speech recognition application.
9     </salt:prompt>
10    <salt:prompt id="askPIN">
11       Please say your password.
12    </salt:prompt>
13    <salt:listen id="recoPIN" onreco="PIN.submit()">
14          <salt:grammar src="PINdigits.grxml" />
15          <salt:bind targetElement="iptPIN" />
16    </salt:listen>
17   </body>
18 </html>
```

3. Példa

X+V

Az XHTML+Voice (X+V) az IBM és az Opera Software által kifejlesztett jelölő nyelv, a VoiceXML mellett az XHTML grafikus képességét használja multimodális alkalmazások fejlesztésére. A SALT-hoz hasonlóan ez a specifikáció is „hangos" jelölőket ágyaz a meglévő XHTML kódba, de nem vezet be újakat, hanem a VoiceXML 2.0 szabványban szereplőket használja. A *<sync>* jelölő segítségével köthetjük a felismert beszédet XHTML változókhoz. Az X+V alkalmazás végrehajtását a VoiceXML űrlapvezérlő (FIA) algoritmusa is szabályozhatja, de a gazdanyelv is gondoskodhat a vezérlésről.

Az X+V és a SALT is nyílt specifikáció és nem hivatalos szabvány, de valószínű, hogy a nyelv néhány eleme bekerül a W3C jövőbeni szabványaiba.

xHMI

Az Extensible Human-Machine Interface (xHMI) a Nuance (régebben Scansoft) által az utóbbi időben meghirdetett nyílt specifikáció, ami kompatibilis a VoiceXML és SALT formanyelvekkel, de a dialógus magasabb szintű vezérlését definiálja. Az xHMI lehetővé teszi a dialógusok közös, nyílt formában történő leírását, mely független a későbbi felhasználás módjától és az alkalmazott technológiától.

4.2. Beszéd bemenet

A beszéd bemenet azokat a funkciókat jelenti, amelyek lehetővé teszik, hogy a felhasználó beszéljen a rendszerhez, a rendszer megértse ezeket a közléseket és megfelelően reagáljon rájuk. A beszéd elemzése a beszédfelismerő feladata. Maga a beszédfelismerés nem standardizált, de szinte minden kereskedelmi beszédfelismerő nyelvtanon alapul, vagy legalábbis a felismerendő egységek formális definícióján.

A W3C Hangböngésző Munkacsoportja a Speech Recognition Grammar Specification (SRGS) jelölőnyelvet definiálta nyelvtanok létrehozására.

SRGS

Az SRGS 2004 óta W3C-ajánlás, nincs konkrét terv a következő verziójára, de ez változhat, ha a piaci szereplők újabb funkciók megvalósításának igényével lépnek fel. Az SRGS két változatban érhető el: XML és ABNF (Augmented Backus-Naur Format). Az ABNF tömörebb, az ember számára jobban olvasható, az XML alapú pedig a gép számára könnyebben feldolgozható. Mivel a nyelvtan definíciója a beszédalapú alkalmazások fejlesztésének legnehezebb része, a szabvány létrejöttének rendkívül nagy jelentősége van az egyéni megoldások használatával szemben. A 2. példa 9-18. sorában egy egyszerű, XML formájú inline („helyben kifejtett") nyelvtanra láthatunk példát.

SISR

Az SRGS kiegészítése a Semantic Interpretation for Speech Recognition (SISR) a W3C új specifikációja. A SISR úgy terjeszti ki az SRGS-t, hogy meghatározhatjuk milyen értékkel térjen vissza a nyelvtan, amikor egy felhasználói közlést felismer. Például bizonyos szituációban az „igen”, „jó”, „oké”, „ja”, „aha” közlések felismeréséhez egységesen azok jelentését az „igen” értéket tudjuk rendelni. A 2. példa 13-15. sorában a szemantikus információ jelölésére használatos *<tag>* elemre láthatunk egy példát.

A SISR „előzetes javaslatterv” állapotban van, a technikai részletek kidolgozottak, de még végső felülvizsgálatra és implementációkra van szükség az ajánlássá válásához.

4.3. Beszéd kimenet

A beszéd kimenet a rendszer által kimondott beszédre vonatkozik. A beszéd kimenet alapulhat szöveg-beszéd átalakítón (Text-to-Speech, TTS) vagy előre felvett beszéd lejátszásán.

SSML

A szöveg-beszéd átalakító bemenete lehet egyszerű szöveg, de gyakran kívánatos jelöltté tenni a szöveget, hogy a beszéd nyelvét, sebességét, a hangsúlyt, a hangerőt, a hangmagasságot, a beszélőt és egyéb tényezőket szabályozhassuk a generált beszédben. Az SSML (Speech Synthesis Markup Languages) biztosítja ezt a lehetőséget. Az SSML egy W3C-ajánlás, amit a W3C Hangböngésző Munkacsoportja fejlesztett ki. Az SSML támogatása követelmény a VoiceXML és a SALT platform számára is.

Kiejtési szótár

A kiejtési szótár (pronunciation lexicon) létrehozása a W3C újabb kezdeményezése, melynek célja, hogy szabványosítsák a szokatlan szavak kiejtését, mind a beszédfelismerő, mind a TTS rendszerek számára. A munka „utolsó felhívás munkatervre” fázisba lépett 2006 januárjában.

4.4. Beszélő azonosítás

A beszélő azonosítás azokat a technológiákat jelenti, amelyek eldöntik, ki a beszélő. Habár jelenleg kimondottan beszélő személy azonosítására nincs szabvány, a biometria néhány szabványa segítségünkre lehet. A BioAPI általános programfejlesztési felület biometriai alkalmazások fejlesztésére ANSI és ISO szabvány.

A CBEFF (Common Biometric Exchange File Format) biometriai adatok leírására szolgáló szabványos adatstruktúra, az XCBF pedig ennek XML alapú verziója. A VoiceXML 3.0 több más újítás mellett a beszélő azonosítás beépítését is ígéri.

1. Táblázat
Beszédtechnológiai specifikációk ([1] alapján)

Név	Technológia/cél	Felelős szervezet	Állapot	Alternatívák
Dialógus szervezés				
VoiceXML 1.0	Dialógus szervezés	VoiceXML Forum	1999-ben hozták nyilvánosságra, mára a VoiceXML 2.0 vette át a helyét	Egyéni megoldások
VoiceXML 2.0	Dialógus szervezés	W3C VBWG	W3C-ajánlás, 2004	Egyéni megoldások, SALT
VoiceXML 2.1	Dialógus szervezés	W3C VBWG	2004 óta felhívás utolsó munkatervre fázisban van	Egyéni megoldások, SALT
VoiceXML 3.0	Dialógus szervezés	W3C VBWG	Követelmények gyűjtése	-
SALT	Dialógus szervezés	SALT Forum	2002-ben hozták nyilvánosságra, azóta nyílt specifikáció	Egyéni megoldások, VoiceXML
	Mutimodális interakció			Egyéni megoldások, X+V
X+V	Mutimodális interakció	Az IBM, az Opera Software és a Motorola összefogása	2001 óta nyílt specifikáció	Egyéni megoldások, SALT
xHMI	Dialógus szervezés és multimodális interakció	Nuance és partnerei	Bejelentették, de még nem hozták nyilvánosságra	Egyéni megoldások
Bemenő beszéd				
SRGS	Nyelvtanok definiálása beszédfelismerésre	W3C VBWG	2004 óta W3C-ajánlás	JSGF, SAPI, Egyéni megoldások
SISR	Szemantikus értékek beszédfelismerőhöz	W3C VBWG	Előzetes javaslatterv, 2006	Egyéni megoldások, JSGF jelölők, SAPI szemantikus jelölők
Pronunciation Lexicon	Kiejtés reprezentálása	W3C VBWG	Utolsó felhívás munkatervre, 2006	Egyéni megoldások
Kimenő beszéd				
SSML	Szöveg kiejtési módjának leírása	W3C VBWG	2004 óta W3C-ajánlás	JSML, SABLE, Egyéni megoldások
Pronunciation Lexicon	Kiejtés reprezentálása	W3C VBWG	Utolsó felhívás munkatervre, 2006	Egyéni megoldások
Kommunikáció				
EMMA	A felhasználói input reprezentálásának formátuma	W3C MIWG	Munkaterv 2004 óta	NLSML
NLSML	A felhasználói input reprezentálásának formátuma	W3C VBWG	Munkaterv 2000 óta	EMMA
MRCP v2	Szétosztott beszédfunkciók	IETF SpeechSc munkacsoportja	2005 decemberében publikálták az utolsó munkatervet	-
DSR-Aurora	Elosztott beszédfelismerési feladatok	ETSI Aurora	Az 1.1.3-as verziója 2003-ban jelent meg	-

4.5. Kommunikáció

A beszéd alapú alkalmazás legfontosabb részeinek leírásán túl, néhány további szabvány az elkészült komponensek kommunikációját biztosítja. A szabványosított kommunikációs protokollok abban az esetben különösen fontosak, ha a különböző rendszerkomponenseket a hálózat erőforrásain szétosztjuk, vagy ha az egyes rendszerkomponensek különböző szállítótól érkeznek.

EMMA

A W3C Multimodális Interakció Munkacsoportja jelenleg is fejleszti az Extensible Multi-Modal Annotation (EMMA) specifikációt, amely a felhasználótól érkező input szabványos leírása. A bemenet forrása tetszőleges lehet: beszéd, kézírás, látás stb. A beszéd alapú alkalmazások esetében a beszédfelismerők így szabványos szövegekkel térhetnek vissza, ami nagyban segíti ezen komponensek integrációját. Az EMMA hamarosan „utolsó felhívás munkatervre" fázisba kerül.

MRCP

A Media Resource Communication Protocol (MRCP) az IETF fejlesztése. Célja, hogy leválassza a beszédfunkciókat (beszédfelismerés, beszédszintézis és beszélő azonosítás) a saját platformjukról úgy, hogy közben szabványos kommunikációs protokollt ír elő az együttműködésükre. Az MRCP v2 a Natural Language Semantics Markup Language (NLSML) szabványt használja – az EMMA elődjét – a felhasználói input reprezentálására.

DSR – Aurora

Az ETSI által definiált Aurora nevű szabvány a beszéd-felismerési funkciókat szétosztja helyi és távoli folyamatokra. Sok esetben előnyösebb, ha lokálisan is végzünk némi beszéd-felismerési feladatot és csak a köztes eredményt továbbítjuk a szerver felé. Például csökkenthetjük a beszédfelismerés hibáját, mivel kevesebb az esély, hogy zaj kerül a beszédjelbe, illetve kisebb sávszélességgel is megelégedhetünk, mivel nem a teljes beszédjel kerül át a szerverre. Ezt a technológiát főképp mobil alkalmazásokban használják.

5. Összefoglalás

A beszéd területén használt szabványos jelölőnyelvek lefedik a dialógusvezérlés, a beszéd be- és kimenet, valamint a komponensek közötti kommunikáció területét. Alkalmazásuktól eszközeink jobb együttműködését, megbízhatóbb technológiai hátteret, gyorsabb, hatékonyabb fejlesztési folyamatot várunk. Természetesen önmagában a szabványok használata nem biztosítja a jó beszéd alapú alkalmazás létrehozását. De ha alkalmazásuk megfelelő fejlesztési tapasztalattal párosul, és figyelembe vesszük az adott felhasználási terület egyéni adottságait, kivívhatjuk a felhasználók elégedettségét.

A jelölőnyelvek dinamikus fejlődése várhatóan tovább folyik a következő években, a W3C két említett munkacsoportjának a működését 2007-ig újra meghosszabbították. Az egyes nyelvek sikerét sok tényező befolyásolja, de az, hogy mennyire találnak támogatásra az egyes fejlesztőkörnyezetekben, illetve, hogy mennyire nyitottak a nemzetköziesítésre, mindenképp a legmeghatározóbbak. A magyar kutatókra, fejlesztőkre vár, hogy ezen, a természetéből adódóan rendkívül nyelvfüggő területen, a szabványok „honosítását" elvégezzék.

A folyamat elkezdődött. 2002-2003-ban a BME Távközlési és Médiainformatika Tanszékén elkészült az első magyar nyelvű VoiceXML böngésző (a felhasznált komponensek részletezése [6] és [7]-ben található). Az MTA SZTAKI Elosztott Rendszerek Osztálya pedig részt vett az EU által támogatott PublicVoiceXML-projektben, melynek célja az első ingyenes és nyílt forráskódú hangböngésző megvalósítása volt [2].

Irodalom

[1] Dahl, Deborah A.:
Guide to Speech Standards.
Speech Technology Magazine, March/April 2005.

[2] Déri András, Fülöp Csaba, Micsik András:
Telefonos szolgáltatások VoiceXML alapon,
NetworkShop 2003 konferencia,
2003. április 14-17., Pécs

[3] Larson, James A.:
VoiceXML:
Introduction to developing speech applications.
Prentice Hall 2003.

[4] Larson, James A.:
State of Speech Standards.
Speech Technology Magazine, July/August 2003.

[5] Kovács, L., Vásárhelyi, Nóra:
Webhez kapcsolódó szabványosítás Magyarországon.
http://nws.iif.hu/ncd2004/docs/ehu/072.pdf

[6] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G.:
„PROFIVOX – A Hungarian Professional TTS System
for Telecommunications Applications",
International Journal of Speech Technology,
Vol. 3, Numbers 3/4, December 2000, pp.201–216.

[7] Szarvas, M., Fegyó, T., Mihajlik, P., Tatai, P.:
Automatic Recognition of Hungarian: Theory & Practice,
Int. Journal of Speech Technology,
Vol. 3, Numbers 3/4, December 2000, pp.237–251.

[8] SALT Forum,
http://www.saltforum.org/

[9] VoiceXML Forum,
http://www.voicexml.org/

Kísérleti gyógyszerinformációs rendszer beszédmodulokkal

OLASZY GÁBOR, NÉMETH GÉZA, BARTALIS MÁTYÁS, KISS GÉZA, ZAINKÓ CSABA, FEGYÓ TIBOR
BME Távközlési és MédiainformatikaiTanszék (BME TMIT)
{olaszy, nemeth, bartalis, kgeza, zainko, fegyo}@tmit.bme.hu
ÁRVAY GERGELY, SZEPEZDI ZSUZSANNA, TERPLÁNNÉ BALOGH MÁRIA
Országos Gyógyszerészeti Intézet (OGYI), gyogyszervonal@ogyi.hu

Lektorált

Kulcsszavak: Profivox-Med gyógyszerfelolvasó, gyógyszernév-felismerő, latin kivételszótár, szövegértelmezés

A „Gyógyszervonal" elnevezésű információs rendszer segítségével a Magyarországon forgalomban lévő gyógyszerek betegtájékoztatóinak szövegéhez juthat hozzá az állampolgár tértől és időtől függetlenül több csatornán. Az információs rendszer elsődlegesen telefonon keresztül lesz elérhető, és egy korszerű, automatikus, beszédalapú dialógusrendszer segítségével fogja a hívó fél számára a gép felolvasni a kiválasztott gyógyszer betegtájékoztatóját. A rendszer két beszédmodult tartalmaz: a Profivox-Med beszédszintetizátort és egy gyógyszernevek felismerésére specializált, kötött szótáras beszédfelismerőt. Ezen felül WEB- és WAP- interfészen keresztül is hozzáférhetők lesznek az adatok. Ismertetjük a rendszer fő jellemzőit, a beszédmodulok fejlesztésének speciális, orvosi területre vonatkozó részleteit. A fejlesztést a BME Távközlési és Médiainformatikai Tanszéke és az Országos Gyógyszerészeti Intézet közösen végzi. A szolgáltatást az Országos Gyógyszerészeti Intézet fogja üzemeltetni, bevezetése 2006 második félévében várható.

1. Bevezetés

Magyarországon jelenleg körülbelül ötezer törzskönyvezett gyógyszer van, melyek engedélyezését az Országos Gyógyszerészeti Intézet végzi. Évente körülbelül 400 új gyógyszer jelenik meg és hozzávetőlegesen ugyanennyit vonnak ki a forgalomból. A nagy fluktuáció miatt a gyógyszerészek és orvosok számára is nehéz feladat naprakész ismeretekkel rendelkezni. **A kitűzött cél, hogy elérhető legyen bárki számára hely- és időkorlát nélkül a gyógyszerekhez tartozó betegtájékoztató szövege.**

A tervezett rendszer olyan korszerű információs szolgáltatást fog nyújtani, amilyen jelenleg nem áll rendelkezésre sem a szakemberek részére, sem a lakosságnak. Az információs rendszer elsődleges célja, hogy telefonon keresztül elérhető legyen, és egy korszerű, automatikus beszédalapú dialógusrendszer segítségével olvassa fel a hívó fél számára a kiválasztott gyógyszer betegtájékoztatóját. A rendszer két beszédmodult tartalmaz: beszédszintetizátort és beszédfelismerőt. Ezen felül WEB- és WAP-interfészen keresztül is hozzáférhetők lesznek az adatok. A rendszert a BME TMIT és az OGYI közösen fejleszti a GVOP pályázati támogatási rendszer keretében.

A gyógyszerekkel kapcsolatos felvilágosítások szinte bárkit érinthetnek a társadalom tagjai közül. Tény, hogy a legnagyobb célcsoport a legtöbb gyógyszert fogyasztók köre, vagyis az időskorúak (mintegy 3 millió nyugdíjas). A másik vélhető célcsoport a fiatalság, akik használják a WAP-ot, az Internetet. Ők segíthetnek az időseknek, ha megfelelően tájékoztatva vannak a szolgáltatás elérhetőségéről. Fontos célcsoport az orvosok köre is, akik az új gyógyszerekről ilyen módon is tájékozódhatnak. A szolgáltatásnak különös jelentősége van a vak és a látássérült emberek részére, mert ők a hagyományos, dobozba csomagolt tájékoztatót nem tudják elolvasni. A „Gyógyszervonal" szolgáltatást az OGYI fogja üzemeltetni, bevezetése 2006. második félévében várható.

2. Rendszerjellemzők

A rendszer főbb paraméterei a következők:
- 24 órás működés (bármikor hívható),
- többféle információs technológiával érhető el a tájékoztató szövege (telefon, Internet, WAP);
- a telefonvonal fogadó végén beszédfelismerő segíti az érdeklődőt, szóban kommunikálhat a géppel;
- a gyógyszerismertetőt gép mondja el, így ezt akár többször is meg lehet hallgatni;
- a gép precíz: megismételt hívás esetén ugyanazt az információt mondja el, ugyanabban a sorrendben, ugyanazon a hangon;
- Internet, WAP használata esetén szövegben kapja meg az információt az ügyfél;
- szakembereknek is tágabb teret ad a 24 órás szolgáltatási forma;
- az információkérés nem hiúsul meg a vonal foglaltsága miatt (megfelelő számú csatorna üzemeltetése esetén).

A rendszer általános blokkvázlata az *1. ábrán* látható.

A rendszerben minden adatot adatbázisban tárolunk, melyek konzisztenciájáért az Internetes modul szerkesztői része illetve maga az üzembentartó a felelős. Az adatbázisban a gyógyszerek alapvető adatai (neve, törzskönyvi száma, hatáserőssége) mellett a hozzájuk tartozó betegtájékoztatók szövegeit, valamint a betegtájékoztatók szövegeinek szintetizált hullámformáit is tároljuk. Ezeket a szövegeket mondatonként tároljuk, minden mondatot csak egyszer. Ennek előnyei:

– Mivel minden mondatot csak egyszer tárolunk, a hibás hangsúlyozású mondatokat csak egyszer kell az előkészítés (szerkesztés) során korrigálni.
– A beszédszintetizátor továbbfejlesztésekor nem kell attól tartani, hogy egy korábban már jónak ítélt mondatot esetleg a frissítés után elront.
– A telefonos elérés esetében fontos, hogy gyorsan elérhetőek legyenek a bejátszandó hangminták. Megfelelően indexelt adatbázisok esetén gyorsabb az adatbázisból kivenni a kész adatokat, mint valós időben előállítani azokat.

3. A tervezés általános kérdései, problémák, nehézségek

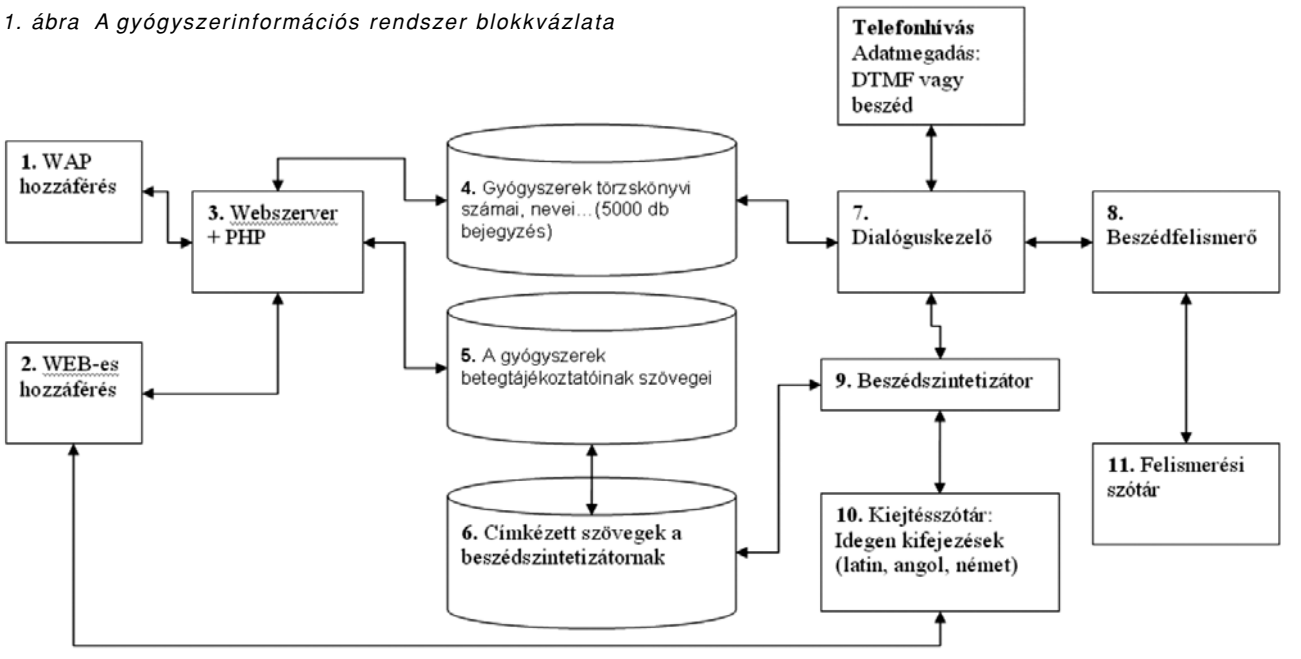
A műszaki és nyelvi fejlesztés több lépcsőben, párhuzamosan zajlik. A fő hangsúly a beszédszintetizátor és beszédfelismerő modulokon van, hiszen ezek fogják biztosítani a hang alapú, párbeszédes üzemmódot. A beszédszintetizátor a gyógyszertájékoztatókat fogja felolvasni, a beszédfelismerőt pedig a hívó fél által kimondott gyógyszernevek helyes felismerésére kell felkészíteni. Egyik sem egyszerű feladat. Ez az első olyan beszédtechnológiai célfejlesztés Magyarországon, amelyik a két módszer kombinálásán túl (párbeszédes alkalmazás) fel van készítve (főleg latin) gyógyszerészeti szakszavak, kifejezések értelmezésére és feldolgozására.

Az előkészítő munka során meghatároztuk azon gyógyszerek listáját, amelyeket kezel a tájékoztató rendszer. Ezek a következők: vény nélküli gyógyszerek, vényre kapható nem kórházi felhasználású gyógyszerek. Minden gyógyszerhez tartozik egy törzskönyvi azonosító szám. Kialakítottuk a gyógyszernevek és a hozzájuk tartozó törzskönyvi számok olyan adatbázisát, mely alapján a keresést el lehet végezni a rendszerben.

A gyógyszerek forgalomba hozatalának engedélyezése hivatalos eljárás, az engedéllyel együtt kiadott alkalmazási előírás és betegtájékoztató hivatalos okiratnak számít. Biztosítani kellett azt, hogy a felolvasandó gyógyszertájékoztató szövege védve legyen az esetleges változtatásoktól, hiszen a rendszerben elektronikus formában, adatbázisokban tároljuk a szövegeket. Pontosan azt kell felolvastatni a géppel, ami a hivatalosan jóváhagyott szöveg. A fejlesztés során szembekerültünk azzal a problémával is, hogy az eddigi betegtájékoztató jóváhagyási ügymenet során nem figyeltek kellő mértékben a szöveg betű szintű helyességére, noha a szöveg minden gyógyszer mellé kinyomtatásra került. A gép a betűk szerint olvas, nem korrigál automatikusan, mint a szem, amikor emberek olvasnak. A helytelen szöveg, az elütések, a helyesírási hibák rontják a beszédszintetizátor hangzását, érthetőségét.

A következő főbb hibacsoportokat állapítottuk meg (példákkal is illusztráljuk):
– betűkimaradás: *mértétől/mértékétől; időponját /időpontját; h a gyógyszer/ha a gyógyszer; zolgál/szolgál*;
– betűbetoldás: *magzatatot/magzatot; Aeurius /Aerius*;
– betűcsere: *Bleocin inejkció/Belocin injekció*;
– helytelen karakter a szövegben (elütés): *1x“ 4 mg-os tabletta*;
– helytelen karakter konverzió: *legfeljebb 25^0C-on/25°C-on*;
– rövidítés helytelen írásmódja, nincs pont: *ill* ;
– mondat a mondatban: - *...kevés folyadékkal (nem grapefruit lével!) étkezés után...*
– idegen szavak többféle írásmódja: *migraine és migrain*
– nem egységes szövegszerkezet: más a logikai sorrend, mivel minden gyár másfajta fogalmazást valósít meg.

1. ábra A gyógyszerinformációs rendszer blokkvázlata



Ezeket a hibákat javítani kell. A korrigálásra olyan korrektúrázó eljárást fejlesztettünk ki, amely nem sérti a hivatalos okirattal szemben támasztott követelményeket.

3.1. A felhasználói felületek tervezése

A felhasználói felületek közül a legbonyolultabb a **telefonos rendszer** működtetését biztosító dialógus („párbeszéd" az ügyfél és a gép között) megtervezése és kialakítása. Ebben biztosítani kell az ember-gép közötti élő párbeszéd optimalizált, mégis kötött formáját. A tervezéskor a legnagyobb problémát a gyógyszerek keresésének, azonosításának egyszerű megvalósítása jelenti (a hívó fél szeretné egy gyógyszerismertető felolvasását kérni, ehhez a gépnek meg kell azt találni a belső adatbázisokban). A gyógyszereket a dialógusban alapvetően két különböző módszerrel azonosíthatjuk, vagy a telefon billentyűzetével bevisszük a gyógyszer valamelyik egyedi adatát, vagy bemondjuk a gyógyszer nevét, amelyből a beszédfelismerő megpróbálja azonosítani a gyógyszert az adatbázisban.

A **nyomógombos bevitelnél** több lehetőség közül választhat a tervező. Az egyik kézenfekvő megoldás, amikor a rendszer a gyógyszer úgynevezett törzsszámának bebillentyűzését kéri a hívó féltől. A gyógyszer törzsszáma egy rövid azonosító, amely egy szöveges résszel kezdődik, majd egy általában 4-5 karakteres számmal fejeződik be. A szöveges rész nem lényeges, a 4-5 karakteres számot könnyű bebillentyűzni a nyomógombokkal. Ez biztos eredményt ad, de az emberek többsége nem ismeri ezeket a számokat, a gyógyszer dobozán sem találhatók meg egyértelműen, valamint a vak és gyengénlátók ezt el sem tudják olvasni, amíg Braille-írással fel nem tüntetik.

Másik megoldás lehet, hogy a felhasználó bebillentyűzheti a gyógyszer nevét is, hasonlóan az SMS íráshoz. Ez főleg idősebb felhasználóknál nem lehet népszerű. Harmadik eset, hogy az ABC betűcsoportjaihoz gombokat rendelünk, például: ABCD=1-es gomb, EFGH=2-es gomb, hasonlóan a T9 bevitelhez. A gyógyszer nevének betűit szóban kéri be a rendszer (például; adja meg a gyógyszer első betűjét a megfelelő gomb megnyomásával). Átlagosan 10-12 billentyűnyomással meghatározható a keresett készítmény.

A bemondáson alapuló megoldáshoz **beszédfelismerőt** kell alkalmazni. Ez természetesebb a felhasználó számára, azonban rendszertechnikailag sok új problémát vet fel. A legnagyobb probléma a gyógyszernevek természetéből adódik, mivel ezek általában latin alapú elnevezések, amelyeknek nincs széles körben elfogadott egységes kiejtése, emellett esetleg több szóból is állhatnak. Az ügyfélnek a gyógyszer nevét kell bemondania a telefonba és a beszédfelismerő azonosítja azt a belső felismerési szótára segítségével. Ez sem egyértelmű feladat, hiszen a gyógyszer neve mellett gyakran szerepel a gyártó neve is (például: Bayer Aspirin), vagy valamilyen hatáserősségre utaló szám (Vitamin C 100 mg filmtabletta), és előre nem lehet tudni, hogy a hívó fél hogy fogja mondani az ilyen neveket. A gyógyszer nevének kiejtési variáltsága is többféle lehet. Fel kell mérni azt, hogy mi lehet az emberektől elvárható kiejtés és több variációra is fel kell készülni.

A fent leírt módszerekkel sok esetben a gép nem tudja egyértelműen azonosítani a gyógyszert, több jelöltet is talál az adatbázisban. A tervezési célunk az, hogy 3-5 lehetséges készítményre lehessen leszűkíteni a keresés eredményét. Ekkor már lehetőség van a készítmények egyenkénti felsorolására, amelyből a felhasználó már kiválasztja azt, amelyikre gondolt.

A gyógyszer kiválasztása után a rendszer felajánlja az ügyfélnek, hogy az adott betegtájékoztató melyik fejezetét *(1. táblázat)* akarja hallani. Az adott fejezeten belül, – miután a gép elkezdte a felolvasást – lehetőség van a mondatok között előre, hátra ugrani, illetve az aktuális mondatot megismételtetni. Ezek az ismétlő technikák teszik jól használhatóvá a rendszert, mivel a felhasználó egy nehezen érthető részt újra meg tud hallgatni, vagy esetleg átugorhatja a számára érdektelen részeket.

1. táblázat A betegtájékoztató hat fejezete

Az eredeti sablon szerint a fejezet címe a doc fájlban	Jelző karaktersorozattal ellátva, gépi szortírozáshoz, kereséshez
1. Milyen típusú gyógyszer X és milyen betegségek esetén alkalmazható?	<<<1>>> 1. Milyen típusú gyógyszer a/az X és milyen betegségek esetén alkalmazható?
2. Tudnivalók az X <szedése> <alkalmazása> előtt	<<<2>>> 2. Tudnivalók az X <szedése> <alkalmazása> előtt
3. Hogyan kell <szedni> <alkalmazni> X-t?	<<<3>>> 3. Hogyan kell <szedni> <alkalmazni> X-t?
4. Lehetséges mellékhatások	<<<4>>> 4. Lehetséges mellékhatások
5. A készítmény tárolása	<<<5>>> 5. A készítmény tárolása
6. További információk	<<<6>>> 6. További információk

A WEB-es és WAP-os felületeknél a kiválasztás és a megjelenítés megvalósítása egyszerűbb, mivel itt billentyűzeten és kijelzőn keresztül történik a kommunikáció az ügyfél és a gépi rendszer között. Akár egy keresőszóra a rendszer által talált 3-5 jelölt közül a felhasználó ki tudja választani a képernyőn, hogy melyik gyógyszerről kéri a tájékoztatót.

3.2. Egységes szövegszerkezet kialakítása

A betegtájékoztatók általában hosszúak, hiszen részletes ismertetést adnak a gyógyszerről. Nem célszerű ezt a szöveget egyhuzamban, az elejétől a végéig felolvasni, mivel ez egyrészről hosszadalmas, másrészről nem biztos, hogy az érdeklődő az egészet akarja hallani.

Az előkészítő munka során megmutatkozott, hogy biztosítani kell a felhasználó részére a választási lehetőséget az egyes fejezetek között. Ezért egységes szövegszerkezeti formát dolgoztunk ki. Fejezetekre osztottuk a szöveget (az OGYI által adott sablon szerint) és minden tájékoztatót ugyanabban az egységes szöveges formában tárolunk az adatbázisban. A sablon szerint a betegtájékoztató elején (bevezetés) általános, fontos információk találhatók, majd a második részben 6 fejezetpont szerepel, 1. , 2. , 3. , 4. , 5. , 6. számozással (lásd az 1. táblázatban). A mintegy 5000 betegtájékoztató szövegét a fenti formára hozva alakítottuk ki a rendszer szöveges adatbázisát (az 1. ábrán az 5. blokk), amelyben a keresés folyik.

Példaként alább bemutatjuk egy ilyen betegtájékoztató 1. pontját:

<<<1>>>

„1. MI A 3TC ÉS MIRE HASZNÁLHATÓ?

A 3TC belsőleges oldat 240 ml oldatot tartalmaz, fehér polietilén flakonba és kartondobozba csomagolva. A csomagban szájon át történő adagolásra szolgáló fecskendőt és a flakonhoz való adaptert is elhelyeztek.

A 3TC az úgynevezett antivirális gyógyszerek egyik csoportjába, a nukleozid analóg reverz transzkriptáz gátlóknak (NRTI) nevezett antiretrovirális szerek közé tartozik. Ezek a gyógyszerek a humán immunhiány vírus (HIV) fertőzés kezelésére szolgálnak.

Az 3TC-t HIV-fertőzött felnőttek és gyermekek kezelésére használják, egyéb antiretrovirális gyógyszerekkel kombinálva. A 3TC csökkenti a HIV vírus mennyiségét az Ön szervezetében és alacsony szinten tartja azt. A CD4 sejtszámot is növeli. A CD4 sejtek egy fajta fehérvérsejtek, melyek fontos szerepet játszanak az egészséges immunrendszer fenntartásában, amely segít leküzdeni a fertőzéseket. A 3TC kezelésre adott válasz betegenként különböző. Orvosa ellenőrizni fogja az Ön kezelésének hatékonyságát."

4. A beszédszintetizátor

A gyógyszerinformációs rendszer beszédszintetizátora a Profivox szövegfelolvasóra alapozott [1] speciális szoftver (BME TMIT fejlesztés), amelyik kifejezetten erre a célfeladatra készült (Profivox-Med). A szoftver specialitását a hagyományos szövegfelolvasókkal szemben két pontban lehet összegezni. Az egyik, hogy érzékeli a latin és egyéb idegen nyelvű szakszavak jelenlétét a szövegben és azokat a magyar kiejtésnek megfelelően olvassa fel. A másik, hogy fel van készítve a gyógyszerészek által használt speciális nyelvezet (mondatszerkesztés, szóhasználat) mondatprozódiai értelmezésére, feldolgozására és megvalósítására a hangsúlyozás, a tagolás, a ritmika, és a beszéddallam tekintetében.

A latin szavak kiejtésére szó szinten dolgoztunk ki betűsorozat – hangsorozat konvertáló szabályokat és ezekkel tulajdonképpen kétnyelvűvé tettük a szintetizátort. Ebben a munkában az OGYI szakemberei voltak segítségünkre. A *2. táblázatban* példát láthatunk a Profivox-Med számára készített kiejtési szabályok gyűjtéséből.

A gyógyszerészek bonyolult nyelvezettel, tömören fogalmaznak a betegtájékoztatóban. Hosszú, összetett mondatokat szerkesztenek felsorolásokkal, gyakori zárójeles betoldásokkal. Legyen példa erre az alábbi négy mondat:

„Amennyiben a parenterális táplálás keretén belül az Aminosteril N-Hepa 8% infúziót egyéb tápanyagokkal (szénhidrátokkal, zsíremulziókkal, elektrolitokkal,

2. táblázat Példa a beszédszintetizátor számára készített kiejtési szabályokból

Eredeti szöveges formátum	A fő gyógyszer neve	magyar kiejtés	szabály	szabály	szabály
ACARBOSE-BAYER	ACARBOSE	akarboze	c>k	s>z	
ACCOLATE	ACCOLAT	akkolat	cc>kk		te#>t#
ACICLOSAN	ACICLOSAN	aciklozán	c>k	s>z	
ACTILYSE	ACTILYSE	aktiliz	c>k	y>i	se#>z#
ADRIBLASTINA PFS/RTU	ADRIBLASTINA	adriblasztina	s>sz		
ADRIBLASTINA RD			s>sz		
AETHOXYSKLEROL	AETHOXYSKLEROL	etoxiszklerol	th>t	s>sz	
ALKA-SELTZER	SELTZER	zelcer	s>z	tz>c	
ALKA-SELTZER N			s>z	tz>c	

vitaminokkal, illetve nyomelemekkel) együtt szükséges alkalmazni, akkor az csak a sterilitás szabályait gondosan betartva, jól összekeverve, és mindenekelőtt a komponensek fiziko-kémiai összeférhetőségére (kompatibilitására) ügyelve történhet.

Segédanyagok: mannit, hidroxi-propil-metil-cellulóz 2910, nátrium citrát, citromsav mononitrát, dinátrium-edetát, tiloxapol, nátrium-hidroxid vagy koncentrált sósav a kémhatás beállítására, tisztított víz.

2 ampulla Alprostapint tartalmát (40 µg PGE [1]) 50-250 ml vivőanyagban feloldva, 2 óra alatt infundáljuk intravénásan.

Ritkán vérelváltozások, vérlemezkeszám csökkenés (thrombocytopenia), májcirrhosisos betegeken eosinphil sejtek számának megemelkedése (eosinophilia), és elszórtan granulocita szám csökkenése (agranulocytosis) is előfordultak."

Erre a nyelvezetre dolgoztunk ki szövegértelmezőt és annak működését percepciós tesztekkel ellenőriztük. A szövegértelmező egyik fontos eleme a felsorolások kezelése. Az élő beszédben az ilyen esetekben a felsorolandó elemek közötti szünetek hosszának kialakítására a beszélő személy az úgynevezett csoportosítási szabályokat alkalmazza. Ezt a bemondó automatikusan végzi, a szünetek hosszát két-három felsorolásos elem kimondása után változtatja, ezzel megtöri a hosszan tartó felsorolás szüneteinek egyhangú menetét. Ahhoz, hogy ilyen szabályokat beépítsünk a szintetizáló rendszerbe, először fel kellett ismertetni a felsorolás helyét és tartamát a szövegben. Ezután a kijelölt szövegegységre alkalmaztuk a csoportosítási szabályokat.

A másik szövegértelmezési specialitás a gyógyszer-tájékoztatókban a gyakori zárójelezés. A zárójeles kifejezéseket a kiejtésben bizonyos szempontból különválasztjuk a szövegtől, betoldást érzékeltetünk. Ennek a mondatprozódiai eszközei a szünettartás és az alapfrekvencia csökkentése. E két elem kombinálásával elértük, hogy a zárójeles részek kiejtése az esetek nagy százalékában közel áll a természetes ejtésben megvalósuló formához. A probléma teljes megoldása azért nem lehetséges, mert ezekben az esetekben érteni is kell a szöveg tartalmát ahhoz, hogy a zárójeles szövegrészt a megfelelő mondatprozódiai elemekkel lássuk el (gyakran például kell szünetet tartani a zárójeles szövegrész után, gyakran pedig nem).

5. A beszédfelismerő

A „gyógyszervonal" beszédalapú telefonos felhasználói felületéhez egy gyógyszernévre optimalizált, nagyméretű kötött szótárból dolgozó beszédfelismerő is tartozik. A felismerő szoftver a felhasználó által a telefonba bemondott gyógyszer nevét ismeri fel, és így azonosítja azt a belső adatbázisban. Az eredményt közli a felhasználóval. A felismerő az alábbi tulajdonságokkal rendelkezik:

– telefonon bemondott gyógyszernevek felismerése elfogadható (min. 90%) pontossággal,
– új gyógyszerek megjelenése esetén az egyszerű bővíthetőség biztosítása,
– a telefonos felhasználói felület menürendszerében való navigáláshoz szükséges parancsszavak és opciók felismerése nagy pontossággal.

A tervezett beszédfelismerő motor beszélőfüggetlen, nyílt szótárral rendelkezik, azaz elvileg tetszőleges szó felismerésére képes (a szó meghatározása után) [2]. Elvileg, mert:

– a szavakat helyesen (a kiejtésnek megfelelően) kell megadni a rendszernek,
– ügyelni kell, hogy nagyon hasonló szavak ne kerüljenek a rendszerbe,
– ha mégis vannak hasonló szavak, dialógus szinten fel kell tudni készülni a tévesztési lehetőségekre,
– a sok gyógyszernév miatt a valós idejű feldolgozás speciális megfontolásokat igényel.

Ezen kritériumok teljesítéséhez az alábbi lépéseket kellett elvégeznünk:

a) Összegyűjtöttük az aktuálisan használt gyógyszernevek listáját az OGYI-tól.

b) Megvizsgáltuk, hogy automatikus módszerekkel ezek a nevek átírhatóak-e, és a nagy számú latin és egyéb idegen eredetű szó miatt arra jutottunk, hogy a magyar nyelvben általában használható automatikus fonetikus átírás itt nem használható.

c) Gyógyszerészek, és az OGYI bevonásával összegyűjtöttük a gyógyszernevek helyes, és általában használt „laikus" kiejtését.

d) Ezek alapján kidolgoztunk egy szabályrendszert, amely többnyire jól meghatározza, hogy egyes betűkombinációkat milyen módon ejtenek ki az emberek az egyes szavakban, például: w>w, ch>c, s>z.

e) A hosszabb gyógyszerneveket fel kellett bontani alap névre és kiegészítésre, mert a rendszer használatakor az emberek általában csak az alap nevet mondják be és ez után lehet rákérdezni a speciális elnevezésre. Például Coldrex-ből öt féle található, és ez nem egyedi.

f) Ezen szabályok alapján elkészíthető a kiejtési szótár, amely előnyös, ha minél több gyakorlatban használt fonetikus átiratot, azaz kiejtési lehetőséget tartalmaz, mert annál biztosabban ismeri fel a rendszer a bemondott gyógyszernevet. Ez a lista részben manuálisan készül el, minden egyes nem magyaros kiejtésű gyógyszer névre külön kell meghatározni az átiratokat. A korábban meghatározott szabályok segítenek a munka félig automatizált megoldásában.

g) A fonetikus átiratok eredménye egy több, mint tízezer elemből álló lista, amit a felismerőnek fel kell ismernie. Mivel a lista igen sok elemet tartalmaz, ezért a kezelésére speciális módszereket kell alkalmazni. A felismerőben alkalmazott nyelvi modell első lépésben a párhuzamos fonetikus átiratokat tartalmazza. Ezt a gráfot lehet optimalizálni súlyozott véges automaták-

kal (WFST) [2]. A nyelvi modellt WFST-ben kell ábrázolni, és a WFST-ken végezhető gráf minimalizáló eljárások segítségével lehet tömöríteni a nyelvi modellt. A tömörítés körülbelül az eredeti méret 50%-ára csökkentette a modellt, valamint a felismerő eljárás sebessége is nagyobb lett.

h) Az eladási statisztikák alapján a gyógyszergyakorisági listát lehet figyelembe venni, és a felismerő (akárcsak az ember) hasonló nevek esetén a gyakoribbra fog dönteni. A nyelvi modell optimalizálása során úgy állítottuk be ezt gyakorisági paramétert, hogy figyelembe tudja venni ezeket a súlyokat is.

i) A felismerési hálózat optimalizálását off-line módon hajtjuk végre, vagyis a hálózatban történő változtatások után minden esetben újra kell fordítani a hálózatot, és újraindítani a felismerőt. Az adott alkalmazásnál ez megengedhető.

6. Az Internetes és a WAP modul

Mind az Internetes, mind a WAP modul kialakításánál azt a szempontot tartottuk szem előtt, hogy a felhasználó minél gyorsabban, illetve minél egyszerűbben megtalálhassa az általa keresett készítményt. Ezt elősegítendő, mindkét modul esetében több szempont alapján kereshetünk (például gyógyszer neve, hatóanyaga vagy a törzskönyvi száma alapján). Az Internetes modul esetében lehetőség van a találatok különböző paraméterek szerinti sorbarendezésére is növekvő és csökkenő módon is. A felhasználó ezekből választhatja ki a kívánt gyógyszer betegtájékoztatóját.

Az internetes modul egyben a rendszer belső, üzembentartói *szerkesztői felületeként* is funkcionál. Szerkesztéshez jelszóval lehet a védett adatbázisokhoz hozzáférni. A szerkesztői interaktív programmal a következők végezhetők el:

- Új készítményekre vonatkozó adatokkal lehet feltölteni az adatbázisokat.
- Korábbi betegtájékoztatókat fel lehet újítani (szöveg kiejtésének karbantartása stb.).
- Forgalomból kivont termékeket törölni lehet a rendszerből.
- A Profivox-Med szövegfelolvasó rendszer által használt kivételszótárt lehet szerkeszteni, bővíteni.
- A rendszerben lévő betegtájékoztatók hangos változatait lehet ellenőrizni (hallgatni), ez alapján az esetleges szöveghibákat meg lehet találni (betűelírás, idegen szavak esetében rossz kiejtés) és korrigálni lehet a kiejtésben ezek hatását, továbbá lehetőség van arra is, hogy az esetleges helytelen hangsúlyozást korrigáljuk.

7. Összefoglalás

A „gyógyszervonal” tájékoztató rendszer az első automatikus, nyilvános informatikai tudakozó a gyógyszerek betegtájékoztatójának szövegével kapcsolatosan Magyarországon. A rendszer több csatornán teszi közzé a gyógyszertájékoztatók szövegét (telefon, WEB, WAP). A legrugalmasabb eleme a telefonos, beszéd alapú dialógusra tervezett modul, amelyik speciális beszédtechnológiai eszközöket tartalmaz (beszédfelismerő a gyógyszerek nevére optimalizálva és beszédszintetizátor a gyógyszertájékoztatók szakszövegére optimalizálva).

A rendszer hasznossága több szempontból indokolt. Számos esetben a téves gyógyszerhasználat megelőzhető. Mint az orvoslás szinte minden területénél, valószínűleg ebben az esetben is jóval hatékonyabb a prevenció, mint az esetlegesen szükséges kórházi kezelés. A rendszer lehetőséget teremt társadalmilag hátrányos helyzetű csoportok (például idősek, sérült emberek, hátrányos helyzetű régióban élők) számára is a betegtájékoztatókban rögzített értékes információk megszerzésére. A közegészségügyi információs ellátottság egyik fontos eleme lehet egy ilyen szolgáltatás.

Köszönetnyilvánítás

A fejlesztést GVOP támogatással valósítjuk meg (szerződésszám: 3.1.1-2004-05-0426).

Irodalom

[1] Olaszy G., Németh G., Olaszi P., Kiss G., Zainkó Cs., Gordos G.: Profivox – a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology, Vol. 3-4. Kluwer Academic Publishers, 2000. pp.201–215.

[2] Mohri M., Pereira F., Riley M.: Weighted Finite-State Transducers in Speech Recognition, In Proc. ISCA Automatic Speech Recognition 2000, pp.97–106.

[3] Fegyó T., Mihajlik P., Szarvas M., Tatai P., Tatai G.: Voxenter – Intelligent Voice Enabled Call Center for Hungarian, Proceedings of Eurospeech 2003, 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, September 1-5, 2003.

Folyamatos, középszótáras, beszédfelismerő rendszer fejlesztési tapasztalatai: kórházi leletező beszédfelismerő

VICSI KLÁRA, VELKEI SZABOLCS, SZASZÁK GYÖRGY, BOROSTYÁN GÁBOR, GORDOS GÉZA

BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
vicsi@tmit.bme.hu

Lektorált

Kulcsszavak: automatikus gépi beszédfelismerés, HMM modellek, n-gram modellek, bigram modellek, perplexitás

A Beszédakusztikai Laboratóriumban kifejlesztésre került egy Windows XP alatt működő, statisztikai elvi alapokra épülő, folyamatos beszédfelismerő fejlesztői környezet (MKBF 1.0), amely alkalmas különböző középszótáras 1000-10.000 szavas szövegek betanítására és felismerésére. Új megoldásokat dolgoztunk ki az akusztikai előfeldolgozásban, a statisztikai modellépítésben valamint fonetikai, fonológiai és morféma nyelvi szinteket vonunk be a felismerési folyamatba. A felismerő a statisztikai alapon működő HMM akusztikai fonémamodellekkel, valamint a statisztikai alapú bigram nyelvi modellekel működik, nem lineáris simítást használva. Vizsgálataink során változtattuk a betanító anyagokat és a szótárkészletet. Kétfajta bigram alappal dolgoztunk: először a hagyományos ragozott szóalakokból építettük fel a bigram mezőket, majd a szóalakokat morfémákra bontottuk, és ezekből a morfémákból építkeztünk. A cikkben a tesztelés eredményeiről, a továbbfejlesztéshez nyert tapasztalatainkról számolunk be. A perplexitási vizsgálatok eredményeinek felhasználásával a felismerési biztonságot 70%-ról 91% fölé növeltük.

Bevezetés

A BME Beszédakusztikai Laboratóriumban kifejlesztett folyamatos beszédfelismerő (MKBF 1.0) optimális működését az akusztikai-fonetikai [4] és nyelvi modellek változtatásával állítottuk be. Természetesen a két szint szétválasztása nem mindig lehetséges, hiszen a tesztfelvételek minősége, zajossága, az artikuláció gondossága, mind befolyásolják a felismerés eredményét, így az nem csak a nyelvi modultól függ.

A felvételek mindegyike – mind a betanításnál, mind a tesztelésnél – 16 kHz-en mintavételezett, 16 biten lineárisan kvantált jel, amely a megfelelő előfeldolgozás után kerül felismerésre.

Az **akusztikai modellek betanítását** az MRBA beszéd adatbázissal végeztük [9].

Végeredményben tehát a fonémaszintű felismerőnk 16 kHz mintavételezésű, 17 Bark frekvenciatérbeli derivált, + 17 időbeni derivált, + 17 időbeni második derivált, + energia bemeneti jelvektor mellett, 4-5 állapotú kvázi-folytonos, 24 lépcsős, rejtett Markov-modellekkel (QCHMM), fonéma alappal dolgozik. Az akusztikai, fonetikai szint optimalizálásáról már korábban beszámoltunk [8].

A **nyelvi betanításhoz** a budapesti SOTE II. sz. Belgyógyászati Klinikájától (2700 lelet) és a szegedi Orvostudományi Egyetemről (6365 lelet) gyűjtött korábbi leletanyag korpuszt használtuk. Ezen szöveg korpusz alapján készítettük el a teljes szóalakszótárat, amely 14 331 szót tartalmaz, a kiejtés szótárat és ezek téma szerint osztott kisebb szótárait, valamint a korpusz alapján morfémaszótárat is készítettünk, amelynek nagysága 6 824 morfémaelem.

Teszteléshez az orvosok által bemondott leletanyagot használtuk: a SOTE II. sz Belgyógyászatán készült, szakorvosok bemondásával, a rendszerhez illeszkedő mintavételi és kvantálási paraméterekkel. Az összes felvételből a férfi orvosok bemondásaiból véletlenszerűen, 5 beszélőtől egyenként 4-4 darab, azaz összesen 20 darab gasztroszkópiás felvételt válogattunk ki tesztelési célokra.

Lényegében a nyelvi modellhez bi-gram modelleket használtunk, de az egyik megoldásban a hagyományos szóalakok (lexémák) az alkotó elemek, a másik megoldásban viszont a morfémák. A morfémabontáshoz a Humor morféma elemzőt használtuk fel [5].

1. Bigram nyelvi modellek

1.1. Az endoszkópiai felismerő nyelvi modelljének leírása

A sokféle szómodell közül az angolszász területeken jól bevált n-gram szómodelleket használtuk a nyelvi szintű felismeréshez. Az n-gram modellek segítségével egy tetszőleges korpuszon minimálisra igyekszünk csökkenteni a perplexítás mértékét, aminek következménye a kevesebb hibázás.

Az n-gram modellekben a nyelvi modellek szószekvenciáik valószínűségének halmazából áll:

$$\hat{P}(w_1, w_2, ..., w_m) \tag{1}$$

A szekvencia valószínűsége ekkor:

$$P(w_1, w_2, ...w_m) = P(w_1)\prod_{i=2}^{m} P(w_i | w_{i-1} ... w_1) \tag{2}$$

A kontextust limitálva:

$$P(w_1, w_2, ...w_m) \cong P(w_1)\prod_{i=2}^{m} P(w_i | w_{i-1} ... w_{i-n+1}) \tag{3}$$

ahol $n>0$ tetszőlegesen választott konstans egész. A nyelv olyan tulajdonságokkal rendelkezik, hogy a folyamat során egy későbbi állapot valószínűsége gyakor-

latilag független a kezdőfeltételektől, így n értékére nem kell nagy *n* értéket használni (tipikus értékek 2-től 6-ig).

A fenti valószínűség ekkor a következőképpen számítható:

$$P\left(w_i \middle| w_{i-1} \dots w_{i-n+1}\right) = \frac{N\left(w_i \dots w_{i-n+1}\right)}{N\left(w_{i-1} \dots w_{i-n+1}\right)} \qquad (4)$$

ahol *N(.)* a megadott szekvencia előfordulásainak száma a tanító szöveganyagban. Ehhez a számításhoz nem kell szegmentált hanganyagot használni, a célra legmegfelelőbbek a nagy írott adatbázisok.

1.2. N-gram modellek simítása

A gyakorlatban a lehetetlen méretű adatbázisok készítése helyett az n-gram modellek statisztikai vizsgálata és különböző módszerekkel történő korrigálását alkalmaztuk[6].

A korrigálásra nemlineáris interpolációt használtunk. Mivel utóbbi esetben lényegesen jobb perplexítás-csökkentés érhető el, ezért a nemlineáris interpolációt használtuk [7] , az *absolute discounting* funkciót. Tekintsük most példaként a bigram esetet, ahol a képlet a következő konkrét alakot ölti (5):

$$\hat{P}\left(w^{(j)} \middle| w^{(i)}\right) = \max\left\{\frac{N\left(w^{(j)}, w^{(i)}\right) - D_i}{N\left(w^{(i)}\right)}, 0\right\} + D_i \frac{|V| - n_0\left(w^{(i)}\right)}{N\left(w^{(i)}\right)} P\left(w^{(j)}\right)$$

($|V|$ a szótár számosságát jelöli, $n_0(w^{(i)})$ pedig azon szavak számát, amelyek egyszer sem követték $w^{(i)}$-t.) A nemlineáris interpoláció esetében a *q(k)* eloszlás súlya arányosan megfelel *(K-*n_0*)*-val, ami azon különböző események száma, amelyek legalább egyszer láthatóak voltak a szöveganyagban. Ez érdekes dologhoz vezet a feltételes valószínűségek modellezésekor: ha a megelőző szót *(predecessor word)* egy, vagy csak néhány szó követi, akkor a simítás 'kisebb' mértékű lesz, mintha sok szó követné azt. Erre utal a nemlineáris kitétel a módszer nevében. Ha *D=1*, akkor az egyszer látott események ugyanúgy kezelődnek az algoritmus által, mint az egyszer sem látottak. Ha alkalmazzuk a Leaving-one-out elvet, akkor nem jelentkezik igazán lényegi különbség a perplexításban, ezért a D értékét a lényegesen egyszerűbben kivitelezhető abszolút modell alapján számítjuk, ahol (6):

$$D_i = \frac{|V| \cdot b}{n_0\left(w^{(i)}\right)}, \text{ ahol } \quad b = \frac{n_1}{n_1 + 2n_2}$$

Itt n_1 és n_2 azon bigramok száma, amelyek pontosan egyszer, illetve kétszer szerepeltek a betanító korpuszban.

Kis betanító anyag esetén $\frac{|V|}{n_0\left(w^{(i)}\right)}$ értéke közelítőleg 1, ezért ilyen korpuszok esetén lehet spórolni a számításokkal és D_i=*b* helyettesítést végezni [6].

2. Nyelvi modell tesztelése, perplexitás vizsgálata

A bigram modellek elkészítéséhez – úgynevezett betanításához – nagy méretű, szöveges, a felismerni kívánt szöveget jól közelítő összetételű és stílusú betanító anyagra van szükség. Esetünkben ez a korábban összegyűjtött és megfelelően feldolgozott (helyesírás ellenőrzés, egységesítés, rövidítések feloldása, fonetikai átírás stb.) leletanyag volt. Ezt a leletanyagot négy csoportra osztottuk az alábbiak szerint:

1. *SOTE II. sz. Belgyógyászatról származó felső endoszkópiás leletanyag (budapesti gasztroszkópia)*
2. *SZTE Belgyógyászatáról származó felső endoszkópiás leletanyag (szegedi gasztroszkópia)*
3. *SOTE II. sz. Belgyógyászatról származó alsó endoszkópiás leletanyag (budapesti kolonoszkópia)*
4. *SZTE Belgyógyászatáról származó alsó endoszkópiás leletanyag (szegedi kolonoszkópia)*

A fenti négy csoportból természetesen lehetőség van kombinált anyagok összeállítására is, amely így nagy mennyiségű betanító anyagot szolgáltathat a bigram nyelvi modellezéshez.

MKBF akusztikai szint betanításait az: MRBA adatbázis férfi bemondásaival végeztük.

2.1.Tesztelési körülmények ismertetése

A tesztelés megkezdése előtt felvetődött az a kérdés hogy a rendelkezésünkre álló betanító anyagok közül melyeket használjuk fel a felismerő nyelvi modelljének a betanítására. Az előzetes mérések alapján *(1. táblázat)* észlelhető, hogy a rendelkezésre álló budapesti és szegedi leletek szókészlete kis mértékben korrelál.

A későbbi vizsgálódások azt is megmutatták, hogy a felismerő tesztelésére kijelölt hanganyagok szótárkészlete újabb szavakat tartalmazott az írásos formá-

		Szegedi colonoscopia		Szegedi gastroscopia	
		szó megvan	szó nincs meg	szó megvan	szó nincs meg
Budapesti colonoscopia	szó megvan	1933	1174	1872	1235
	szó nincs meg	5089	6135	7067	4157
		Szegedi gastroscopia		Szegedi colonoscopia	
		szó megvan	szó nincs meg	szó megvan	szó nincs meg
Budapesti gastroscopia	szó megvan	2720	1594	2065	2249
	szó nincs meg	6219	3798	4957	5060

1. táblázat
Szókészletek összehasonlítása

ban rendelkezésünkre álló budapesti- és szegedi endoszkópos leletekhez képest. A fenti táblázat egyértelműen mutatja, hogy a szegedi és budapesti leletek szóhasználata között milyen nagy eltérés van és a tesztelésre kijelölt annotált felvételek budapesti kórházból származnak. Ennek ellenére a fent említett okok miatt az összesített leletanyaggal való betanítás ígérkezett megfelelőnek.

Az egyes rövidítések számítási módja (a későbbiekben is ezt a jelölésrendszert alkalmazzuk):

A tesztelésnél használatos mérőszámok:

Össz _ref: *a felismerendő egységek száma,*
Helyes: *a jól felismert egységek száma,*
Del: *a törölt ,*
Össz_ rec:*a felismert egységek száma,*
Ins:*a beszúrt egységek száma,*
Subs:*a helyettesített egységek száma*

$$CORR = \frac{Helyes}{Össz\_rec},\ Acc = \frac{Helyes - Ins}{Össz\_rec},\ \mathbf{Wer} = 1 - CORR$$

2.2. Perplexitás alapú WER becslés

A gépi beszédfelismerés felismerési pontosságát a szakirodalomban – a fentiekben leírt – a Word Error Rate *(WER)* indikátorral szokásos jellemezni.

A Word Error Rate egy költséges művelet eredménye, ezért szükségessé vált egy olyan indikátor bevezetése, amely a beszédfelismerés akusztikai szintjétől függetlenül becslést tudna adni a felismerés pontosságára. Így a nagy felismerési idő és a költséges WER számítás kikerülhető lenne a beszédfelismerés nyelvi modelljének vizsgálata esetén. Egy ilyen becslési módszer a – szakirodalomban is jól ismert – perplexitás, melynek segítségével vizsgálni tudjuk a nyelvi modellt.

Bár különböző kutatások rávilágítanak hogy készíthető paraméterfüggő (betanítás, nyelvi modell, akusztikai modell) becslési eljárás [3,4], mégis a konkrét paraméterek ismeretének hiányában a perplexitást találtuk olyan becslési eljárásnak amely a szakirodalomban elfogadott és számítási módja ismert. A perplexitás számítási módját az alábbi képletben ismertetjük:

$$PP = \frac{1}{\left(\prod_{i=1}^{N} P(W_i \mid W_{i-1})\right)^{\frac{1}{N}}} \qquad (7)$$

ahol W_i teszt szekvencia *i.* **szava**; W_{i-1} a teszt szekvencia *i*-1. **szava**, N a teszt szekvenciát alkotó **szavak** száma.

A perplexitás képletét bigram alapú nyelvi modell formájában használtuk. A **szavak** jelentése lexéma szintű felismerés esetén lexéma, míg morféma alapú felismerés esetén morféma.

A perplexitás értékkészlete egy 1-nél nagyobb valós szám. A tesztanyag nyelvi modul általi felismerése annál tökéletesebb, minél jobban közelít a perplexitás értéke 1-hez. Minél nagyobb a perplexitás értéke, annál kevésbé fedi a betanítás a tesztelő szövegkörnyezetét.

3. A tesztelési eredmények ismertetése

A következő táblázatokban a felismerő tesztelési eredményei láthatóak:

Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
1173	1580	750	451	22	401	25,4	36,1

2. táblázat
Gasztroszkópiás felvételek tesztelési eredményei lexéma alapú összegzett betanítású nyelvi modell esetén, orvosok bemondásában

Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
890	1326	504	822	8	370	35,7	43,4

3. táblázat
Colonoszkópiás felvételek tesztelési eredményei lexéma alapú összegzett betanítású nyelvi modell esetén, orvosok bemondásában

A viszonylag rossz eredmények oka (tipikusan a kötőszavak tévesztése nagy), hogy bár a szótárkészlet ezen betanítóanyag választása mellett biztosítja a legnagyobb fedést, ennek ellenére bigram szókapcsolatok nem fedik a tesztelési bigram szókapcsolatokat. Ha megfigyeljük az 1. táblázatbeli eredményeket, és összevetjük a tapasztaltakkal, akkor megállapíthatjuk, hogy a teljes anyaggal történő betanításkor (szegedi, budapesti, gasztroszkópiai, colonoszkópiai) olyan nagy mértékű hamis szókapcsolat-statisztikát vittünk be a rendszerbe, hogy az a bigram valószínűségi mezőben igen nagy zaj keletkezett, így bár hiába lettek betanítva ezen szókapcsolatok, mégis rossz lett a felismerés (lásd az 1 . táblázat szókészlet eltéréseit.)

Szűkítve a betanítási anyagot, és csak a budapesti gasztroszkópiás betanítóanyagot használva az eredmények javulnak, amint azt a *4. táblázatban* mutatjuk:

Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1150	1417	799	283	8	343	44,8	30,5	*73,59*

4. táblázat
Gasztroszkópiás felvételek tesztelési eredményei lexéma alapú budapesti gasztroszkópiás betanítású nyelvi modell esetén, orvosok bemondásában

A fenti táblázatok eredményei orvosok bemondásai alapján elkészített tesztelési eredmények. A lelet-felvételek meghallgatásakor azt tapasztaltuk, hogy a felvételek igen zajosak és kiejtés szempontjából is igen rossz minőségűek. Így felvetődött azon lehetőség – a nagyobb felismerési pontosság elérése érdekében – hogy limitált szintű zajkörnyezetben felvett felvételekkel teszteljünk és a felvétel során ügyeljünk a helyes artikulációra. Ennek érdekében a budapesti gasztroszkópiás leleteket – amelyeket az orvosok is bemondtak (20 lelet) – bemondtuk a laboratóriumban, a fentiekben megfogalmazott kritériumok teljesítése mellett és ezen felvételeket használtuk fel az MKBF tesztelésére. A tesztelési eredményeit az *5. táblázatban* közöljük.

Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1173	1451	922	280	1	250	54,7	21,3	*73,59*

5.táblázat
Kiszajú,laboratóriumi gasztroszkópiás felvételek, tesztelési eredményei lexéma alapú budapesti gasztroszkópiás betanítású nyelvi modell esetén

Az eredményeket összevetve a 4. táblázatéval, látható hogy a szótévesztési arány (WER) javult, mivel az akusztikai szintű felismerés javult, ezáltal az egész felismerés is pontosabbá vált.

Elvégeztünk továbbá egy olyan speciális tesztelést, hogy olyan leletbemondással tesztelünk, ami szerepelt a nyelvi modell betanításában. Ennek megvizsgálása érdekében – a fentiekben említett akusztikai feltételek biztosítása mellett – budapesti gasztroszkópiás leleteket rögzítettünk, amelyeket a budapesti gasztroszkopiás írott leletanyagból olvastunk fel, és a nyelvi modellt a 5. táblázathoz hasonlóan budapesti gasztroszkópiás endoszkópos írott leletanyaggal tanítottuk be. Az eredményeket a *6. táblázatban* adjuk meg.

Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER	PP
416	444	380	28	0	36	84,6	8,6	9,36

6. táblázat
Laboratóriumi, budapesti, gasztroszkópiás felvételek tesztelési eredményei lexéma alapú, budapesti, gasztroszkópiás, betanítású nyelvi modellekkel

A táblázat egyértelműen mutatja, hogy a szótévesztési arány (WER) javul, hiszen olyan leletanyagot teszteltünk, ami a betanításban szerepelt. Ha a perplexitást vizsgáljuk, akkor 4. és 5. táblázatokban található perplexitás-értékekhez képest nagy arányú perplexitás csökkenést tapasztalhatunk.

3.1. A lexéma alapú tesztelési eredmények kiértékelése

Az eredmények alapján az alábbi következtetéseket vonhatjuk le:

1. A budapesti és a szegedi leletanyag annak ellenére hogy mind a két korpusz azonos témájú, azonban szóhasználatban, stílusban jelentősen különböznek. Igy e vegyes anyag alapján készített bigram nyelvi modellel, noha ez elviekben robosztusabb, a gyakorlatban mégis gyengébb felismerés érhető el. A *WER* eredmény 5,54%-kal rosszabb a csak budapesti anyag alapján tanított bigramhoz képest.

2. A beszédfelismerés során különösen fontos a szöveg gondos, folyamatos bemondása, így a *WER* értéke akár 10%-kal is lejjebb szorítható.

3. A jelenlegi, budapesti gasztroszkópiás leletanyag alapján készült bigram nyelvi modell nem fedi kellőképpen a kívánt alkalmazási területet. Ezt igazolja a szakorvosok és a saját bemondásban készült leletek hibaarányainak nagyfokú korrelációja, azaz a hibaarány jelentős része a bigram nem megfelelő fedéséből, és nem az akusztikai jel minőségéből adódik.

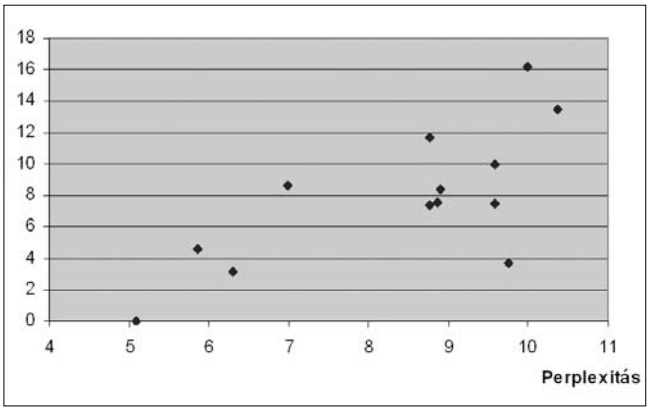
4. Megjegyezzük, hogy a szakorvosok által végzett bemondásokban a PI kódjelű bemondó halk, az általánosan elvárhatónál gyengébb beszédproduktumot adott. Amennyiben az ezekkel a felvételekkel kapott hibaarányt figyelmen kívül hagyjuk, az összesített *WER* értéke 30,52%-ról 24,27%-ra javul.

5. Az *Acc* paraméter esetenkénti alacsony értéke arra enged következtetni, hogy az adott tesztfelvétel a nyelvi modell számára ismeretlen, vagy a bigram betanító anyagában nem kellő gyakorisággal előfordult szót tartalmaz. Ilyenkor a beszúrások megszaporodnak, amely jellemzően több rövid, felolvasva az elhangzó, de fel nem ismert szóéhoz hasonló hangélményt ad.

3.2. A perplexitás vizsgálata és ennek összevetése a lexéma, valamint morféma alapú beszédfelismerő tesztelési eredményeivel

Ebben a részben azt vizsgáltuk, hogy a felismerés hatékonysága menynyire jósolható a nyelvi modell szimulálására szolgáló perplexitás-számítással.

Ennek vizsgálata érdekében megmértük, hogy miként konvergálnak a perplexitás alapú becslési értékek a Word Error Rate (WER) értékekkel abban az esetben, amikor olyan anyaggal tesztelünk, ami részét képezi a nyelvi modell betanításának. A budapesti gasztroszkópiás leletekkel kapott mérés eredményeit *1. ábrán* a szemléltetjük.



1. ábra
Perplexitás és WER értékek korrelációja kiszajú, laboratóriumi, budapesti gasztroszkópiás felvételek esetén, budapesti gasztroszkópia nyelvi modellel tanítva

Amint az ábra eredményei szemléltetik, a perplexitás – WER értékpárosok által meghatározott pontok szórnak, de az összefüggés a perplexitás és a WER értékek között jól látható [1]. Az, hogy az értékek szórnak, érthető, hiszen nem a teljes magyar nyelvet vizsgáltuk, hanem csak egy igen szűk tématerületet, ami – a tapasztalatok alapján – a szakterületi mellett hétköznapi, valamint irodalmi nyelvet is tartalmaz. A másik lehetséges oka abban rejlik, hogy a perplexitásszámítás folyamán nem számolunk a fonématévesztéssel és a fonématévesztés hatására eltolódik a WER-Perplexitás kapcsolat [4].

3.3. Lexéma vagy morféma felismerés

A bigram statisztikákat egyszer lexéma, egyszer morféma alapokon számítottuk. A tesztelések hasonló eredményekre vezettek mindkét esetben.

Ami a morféma alapú felismerés mellett szól:

Ha a morféma szintű felismerést választjuk a szótárméret jelentősen csökken, így kisebb bigram valószínűségi mezőt kell kezelni. Lexéma alapú betanítás esetén a leletkorpusz alapján 14331 (lexéma) egység jön létre, míg morféma alapú esetén 6706 egység (morféma).

Mivel a bigram valószínűség mező egy négyzetes diszkrét valószínűségi mező, így tárolási szempontból körülbelül 4,5-szeres tárcsökkenést eredményez, és még a legrosszabb esetben (simítás esetén) is már átlagosan 2,13-szeres valószínűségérték növekedés érhető el.

Morféma alap								
Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1631	2045	1241	778	9	355	28,3	23,9	*27,31*
Lexéma alap								
Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1173	1451	922	280	1	250	54,7	21,3	*73,59*

8. táblázat
Tesztelési eredmények, betanítás a budapesti gasztroszkópiás anyaggal, tesztelés kiszajú, laboratóriumi, budapesten felvett gasztroszkópiás felvételekkel

Ami a lexéma szintű felismerés mellett szól:

Morféma szintű felismerés esetén komoly problémát jelent a toldalékok határain fellépő hasonulások, összeolvadások, hangrendilleszkedések hangzókiesések kezelése. Ennek leírása egyelőre úgy tűnik, csak manuálisan oldható meg.

4. Felismerési pontosság növelése perplexitás alapú szimulálás segítségével

Amint azt az 1. fejezetben ismertettük, a perplexitással becsülni lehet a felismerés pontosságát.

A 5. táblázat egy olyan tesztelési eredményeket tartalmazó táblázat, ahol a tesztelésnek kinevezett állomány nem szerepelt a nyelvi modell betanító leletei között. A 6. táblázat viszont olyan tesztelési eredményeket tartalmaz ahol a tesztanyag részét képezte a nyelvi modell betanító anyagának, tehát biztosítottak voltak azon szókapcsolatok betanításai amelyek a tesztelésnek kinevezett anyagokban szerepeltek. Ha a 5. táblázat eredményeit összszevetjük a 6. táblázat eredményeivel, azt tapasztalhatjuk hogy a hibaszázalékok kisebbek a 6. táblázatban.

Így felvetődött az a kérdés, hogy ha betanítanánk a 5. táblázathoz tartozó tesztelési mondatokból azon szókapcsolatokat amelyek nem szerepeltek a betanításnál, akkor a felismerési pontosság várhatóan növekedni fog-e. Ehhez csupán ezen hiányzó szókapcsolatokat kell megkeresni és a betanítóanyagban elhelyezni.

A hiányzó szókapcsolatok betanításánál a tesztelő anyagból azon **szóláncokat** kerestük meg, amely szóláncok bármely bigram szókapcsolatát tekintve sem szerepeltek a betanításban. Ez formálisan tehát a következő:

<utolsó betanításban szereplő szó>
<betanításban nem szereplő szó>
+ <első olyan szó ami a betanításban szerepelt>

a + jel jelenti, hogy 1-nél többször is szerepelhet egymás után betanításban nem szereplő szó, a reguláris kifejezéseknél használt jelöléssekhez hasonlóan

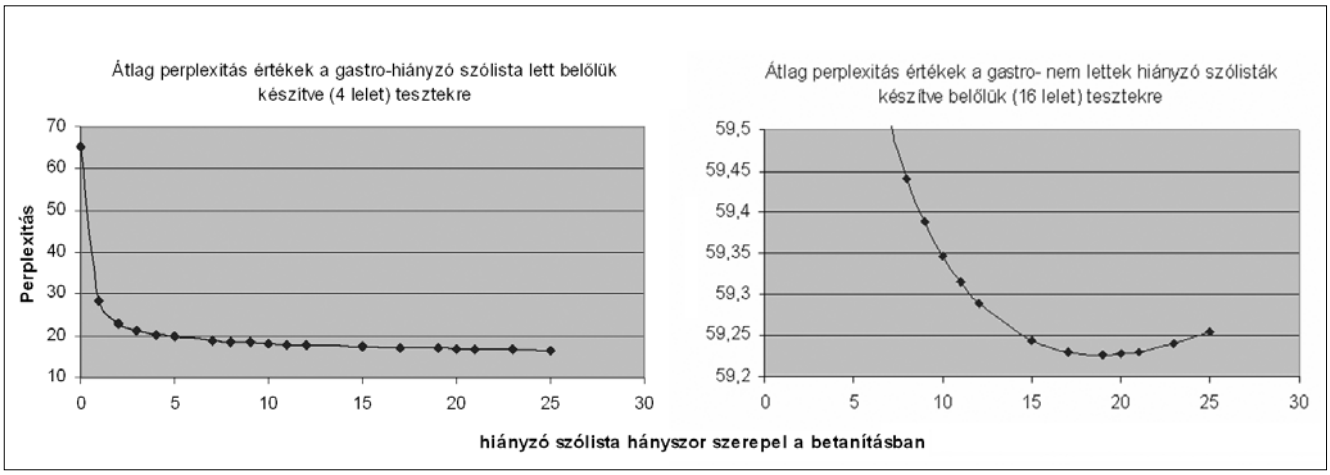
Ezen módszer választása mellett a szándékunk az, hogy a hiányzó bigram valószínűségeket betanítsuk anélkül hogy a már meglévő bigram valószínűségeket jelentősen torzítanánk. (Ezáltal azon bigram valószínűségeket is betanítjuk, amelyek a már meglévő bigram valószínűségmezőbe beillesztik ezen új betanításokat. Ezek a szóláncok első és utolsó szavai, hiszen így a hiányzó szóláncnak a környezetét is be tudtuk tanítani a hiányzó szókapcsolatok mellett.)

Azt tudjuk, hogy mely részekkel kell a betanítást bővíteni, azonban azt nem, hogy ezen hiányzó részek betanítását hányszor kell megismételni. Ezen szám meghatározása végett arra az összefüggésre építünk, hogy perplexitással megbecsülhető a nyelvi modell felismerési pontossága. A mérés ezen a perplexitás alapú nyelvi modell hatékonyságának becslése alapján történt. Előállítottunk különböző betanító anyagokat, melyek felépítésüket tekintve a következőképpen alakultak:

Betanítás=
<budapesti gasztroszkópiás leletek>
+ <hiányzó **szóláncok** >*

*a * jel jelenti, hogy a hiányzó szóláncok 0...n-szer szerepelhetnek a budapesti gasztroszkópiás leletek után, a reguláris kifejezéseknél használt jelöléssekhez hasonlóan.*

2/a. és 2/b. ábra Perplexitás átlagok alakulása



A perplexitás értékeket különböző betanító anyagoknál nem lehet összehasonlítani, esetünkben viszont az összehasonlítás elvégezhető, mivel a betanítóanyagot csak kismértékben módosítottuk, szókészletek megegyeznek, a bigram valószínűségi mező csupán eloszlási értékeiben csak kismértékben különbözik egymástól.

4.1. A betanításszám meghatározása nyelvi modell szimulálás segítségével

Kiválasztottunk négy leletet tesztelésre, a betanítóanyaggal összehasonlítva meghatároztuk a hiányzó szóláncokat. A betanításnál ezen hiányzó szóláncokat szerepeltettük 0...n-szer (3, 33, 53, 92 leletek).

Ezen leletek mellett figyeltük a többi leletet is, hiszen a cél a bigram valószínűség mező felismerésének erősítése, nem pedig a torzítása, hiszen ezen hiányzó szóláncok egy határszint feletti ismétléses betanítása azt eredményezné, hogy e négy leletet ismerné csupán fel a felismerő, márpedig nem ez a célunk. A célunk az, hogy e négy gasztroszkópiás leletre a felismerési pontosság javuljon, annak figyelembevétele mellett hogy a többi tesztanyag – amelyekből nem lett hiányzó **szólánc** kiválasztva – tesztelési eredményei ne romoljanak le markánsan. A mérési eredmények grafikusan a *2. ábracsoporton* láthatóak.

Amint a *2/a.* ábrából megfigyelhető a hiányzó szóllánc ismétlési számának növelésével a perplexitás értékek javulnak azon tesztanyag esetén, amely alapján a hiányzó szólánc elő lett állítva. Az is látható a 2/a. ábrából, hogy már egyszeri betanítás hatására is milyen jó perplexitás javulást érhetünk el. A *2/b.* ábrán szemléltettem azon gasztroszkópiás leletek perplexitás értékeinek alakulását, amelyekből nem lett hiányzó szólánc véve. Látható, hogy a 19-20-szoros betanításig folyamatos perplexitás javulás van, e feletti betanításnál viszont csak romlás tapasztalható.

Mindezeket figyelembe tartva a 20-szoros betanítás mellett döntöttünk, hiszen a hiányzó szólánc kigyűjtése azon gasztroszkópiás tesztekre hatott a leginkább, melyekből nem lett hiányzó szólánc készítve (2/b. ábra).

A hússzoros betanítással kiegészített budapesti gasztroszkópiás anyaggal újra elvégeztük a tesztelést:

Betanítás:
1: budapesti gasztroszkópiás írott leletanyagok
+<különbözeti szólánc>
2: budapesti gasztroszkópiás írott leletanyagok
*+ 20*hiányzó szólánc*
+<különbözeti szólánc>

Tesztelés:
3, 33, 53, 92 leletek,
amelyekből a hiányzó szólánc készült.

A 2/a. ábrát figyelembe véve arra számíthatunk, hogy a 3, 33, 53, 92-es leletek felismerése javul. A következő tesztelési eredményeket kaptuk a kétfajta betanítás mellett:

A *9/a. táblázat* az első betanítás tesztelési eredményeit szemlélteti, míg a *9/b. táblázat* a második betanítás eredményeit szemlélteti. Tehát amint azt a táblázatok is mutatják a szóláncok 20-szoros megismétlésével a szótévesztés erősen lecsökkent (9/b. táblázat), a **19,7%-os szótévesztés 8,9 %-ra javult**.

Azonban kérdéses, hogy a tesztelésnél megjelölt többi lelet esetében mi lett az ilyen betanítás mellett az eredmény. A perplexitásos vizsgálatok alapján ezen teszt leletek felismerése nem romolhat. A következőkben tehát azon tesztelési eredményeket mutatjuk be, amely tesztanyagokból nem lett hiányzó szólánc készítve. Ezen eredményeket tűntetjük fel a *10/a. és10/b. táblázatokban*.

9/a. táblázat
Tesztelési eredmények
budapesti gasztroszkópiás írott leletanyagok esetén.
Tesztanyag: 3, 33, 53, 92

betanítás: hiányzó szólánc nem szerepel a betanításban									
Lelet	Bemondó	Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
3	BG	95	116	67	21	0	28	48,4	**29,4**
33	BG	63	80	48	17	0	15	49,2	**23,8**
53	SG	55	68	50	13	0	5	67,2	**9,1**
92	SG	55	62	46	7	0	9	70,9	**16,3**
átlagos_WER:		**19,6%**	átlagos Acc:		**58,9%**				

9/b. táblázat
Tesztelési eredmények
budapesti gasztroszkópiás írott leletanyagok
*+ 20*hiányzó szólánc esetén.*
Tesztanyag: 3, 33, 53, 92

betanítás: hiányzó szólánc 20-szor szerepel a betanításban									
Lelet	Bemondó	Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
3	BG	95	110	78	17	1	16	64,2	**17,8**
33	BG	63	69	61	6	0	2	87,3	**3,1**
53	SG	55	58	53	3	0	2	90,9	**3,6**
92	SG	55	61	49	6	0	6	78,1	**10,9**
átlagos_WER:		**8,9%**	átlagos Acc:		**80,1%**				

10/a. táblázat
Tesztelési eredmények
budapesti gasztroszkópiás írott leletanyagok esetén.
Tesztanyag: 94, 38, 174

betanítás: hiányzó szólánc nem szerepel a betanításban									
Lelet	Bemondó	Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
94	SG	51	64	41	13	0	10	54,9	**19,6**
38	ZT	34	41	11	11	2	21	0	**67,6**
174	SG	118	167	70	49	0	48	17,7	**40,6**
átlagos_WER:		**42,6%**	átlagos Acc:		**24,2%**				

10/b. táblázat
Tesztelési eredmények
budapesti gasztroszkópiás írott leletanyagok
*+ 20*hiányzó szólánc esetén.*
Tesztanyag: 94, 38, 174

betanítás: hiányzó szólánc 20-szor szerepel a betanításban									
Lelet	Bemondó	Össz. ref	Össz. rec	Helyes	Ins	Del	Subs	Acc	WER
94	SG	51	64	41	13	0	10	54,9	**19,6**
38	ZT	34	41	11	11	2	21	0	**67,6**
174	SG	118	166	68	48	0	50	16,9	**42,3**
átlagos_WER:		**43,2%**	átlagos Acc:		**23,9%**				

Az eredmények elemzése alapján azon gasztroszkópiás leletek esetén, amelyeknél nem készítettünk szóláncot, javulást várhatunk, de romlásra nem kell számítanunk (2/b. ábra). Ha megtekintjük a 9/a. és 9/b. táblázat 94-es leletének sorát, láthatjuk hogy előzetes becslésünk beigazolódott, hiszen a felismerés nem romlott, hanem ugyanolyan maradt.

A 10/a. és 10/b. táblázatokat osszehasonlítva láthatjuk, hogy nem változott a felismerés pontossága, tehát itt is beigazolódott az előzetes becslés. A 174-as lelethez tartozó sorban egy kismértékű felismerési pontosság romlása tapasztalható, de a többi esetben a pontosság változatlan.

Tehát a perplexitás elemzéssel, és a betanító anyag egyszerű módosításával egy meghatározott szótárkészletű, nyelvi szöveg felismerését jelentős mértékben javítani tudjuk.

5. Összefoglalás

Tehát a 2/b. ábrán példaként bemutatott perplexitás átlag vizsgálata alapján sikerült a szóláncok ismétlési számát optimálisra beállítani, úgy , hogy a felismerés lényegesen jobb lett, 90% fölötti. A vázolt eljárás sok esetben javíthatja a felismerést, nem csak kórházi leletezés esetében, hanem bármely behatárolt témában.

Figyelembe kell azonban venni, hogy a módszerünk nem ad valóságos megoldást, hiszen a gyakorlatban, az erősen agglutináló nyelveknél tisztán statisztikai n-gram modellel dolgozva, mindig lehet új elem az új bemondások között, ami a betanító anyagban nem szerepelt, és ez hibázáshoz vezet. Azonban, egy közepes szótár méretű, kötött témában kialakítandó felismerő létrehozásában jelentős segítség lehet.

Irodalom

[1] Máté Szarvas, Sadaoki Furui:
Evaluation of the Stoachic Morphosyntactic Language Model on a One Million Word Hungarian Dictation Task. EUROSPEECH, Genova 2003. pp.2297–2300.

[2] Stanley Chen, Dougles Beeferman, Ronald Rosenfeld:
Evaluation Metrics For Language Models, In: DARPA98, National Institue of Standards and Technology (NIST), www.nist.gov/speech/publications/darpa98/html/lm30/lm30.htm

[3] Philip Clarkson, Tony Robinson:
Towards improved language model evaluation measures
http://Citeseer.ist.psu.edu/clarkson99toward.html

[4] Yonggang Deng, Milind Mahajan, Alex Acero:
Estimating Speech Recaognition Error Rate without Acoustic Test Data.
http://research.microsoft.com/srg/papers/2003-milindm-eurospeech.pdf

[5] HUMOR Morfológiai elemző.
http://www.morphologic.hu/h_humor.htm

[6] Claudio Becchetti, Lucio Prina Ricotti:
Speech Recognition,Theory and C++ implementation.
Fondazione Ugo Bordoni, Rome, 1999.
ISBN 0-471-97730-6

[7] Ney, H., Essen, U., Kneser, R.:
On Structuring Probabilistic Dependencies in Stochastic Language Modeling.
Computer Speech and Language, 1994/8. pp.1–38.

[8] Velkei Szabolcs, Vicsi Klára:
Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából,
MSZNY 2004., pp.307–315.

[9] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László:
Beszédadatbázis irodai számítógép-felhasználói környezetben,
II. Magyar Számítógépes Nyelvészet Konf., 2004., pp.315–319.

Generációváltás a beszédszintézisben

FÉK MÁRK, PESTI PÉTER, NÉMETH GÉZA, ZAINKÓ CSABA

{fek,nemeth, zainko}@tmit.bme.hu, pesti@alpha.tmit.bme.hu
BME Távközlési és Médiainformatikai Tanszék

Lektorált

Kulcsszavak: formánsszintetizátor, elemösszefűzés, elemkiválasztás, korpusz alapú beszédszintézis, szubjektív minősítés

A cikk áttekinti a beszédszintézis rendszerek három generációjának fejlődését. Bemutatjuk a BME TMIT-en fejlesztett magyar nyelvű kísérleti korpusz alapú, elemkiválasztásos beszédszintetizátor felépítését. Részletesen ismertetjük a hang és szóhatárok automatikus jelölésének módszerét. Feltárjuk a prozódia megvalósításának lehetséges módszereit egy korpusz alapú, elemkiválasztásos beszédszintetizátorban. Ismertetjük az elemkiválasztás működését a fejlesztés alatt álló rendszerben. A kísérleti rendszer beszédminőségét összehasonlítjuk a korábbi magyar nyelvű beszédszintézis rendszerek minőségével.

1. Bevezetés

A beszédszintézis rendszerek célja a bemeneti információ beszéddé alakítása. A bemenet legtöbbször egy felolvasandó szöveg (ekkor szövegfelolvasó (text-to-speech) rendszerről van szó), de lehet valamilyen adat (számlaegyenleg, járatinformáció, időjárási adatok stb.). Az adatjellegű bemenetet kezelő, információ-felolvasó (concept-to-speech) rendszerek első lépésként általában szöveggé alakítják a bemenetet. Egy szövegfelolvasó rendszer két alapvető részből épül fel. Az első rész a bemeneti szöveget alakítja szimbolikus információvá, a második a szimbolikus információt alakítja a beszédjelet leíró hullámformává (általában valamilyen hangfájlt állít elő). A közbenső szimbolikus információ általában a szöveg tartalmát megadó fonéma sorozatból (egy fonéma egy beszédhangot jelöl) és a beszéd prozódiai jellemzőit (hanglejtés, hangsúlyok, ritmika) leíró információkból áll össze.

A cikk első részében áttekintjük a beszédszintézis rendszerek három generációjának fejlődését. A második rész a BME-TMIT-en fejlesztett kísérleti korpusz alapú, elemkiválasztásos beszédszintetizátor működését mutatja be. A befejező részben a három bemutatott generációnak megfelelő három magyar nyelvű beszédszintetizátort hasonlítjuk össze egy szubjektív minősítési teszt segítségével.

2. Formánsszintézis

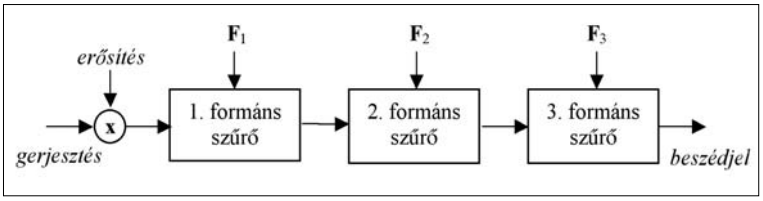
A formánsszintézis volt az első olyan beszédszintézis technológia, melynek segítségével egy szöveget automatikusan folyamatos és jól érthető beszéddé lehetett alakítani. Az elnevezés a szövegfelolvasó rendszerben alkalmazott hullámforma előállítás módszerére utal, ami egy gerjesztett szűrőrendszer kimeneteként állítja elő a beszédjelet. A formánsszintetizátor egy lehetséges megvalósítását az *1. ábra* mutatja.

A formánsszintetizátor az emberi beszédkeltést modellezi. Számítástechnikai erőforrásigénye kicsi. A gerjesztés a hangszalagok által keltett jelnek feleltethető meg: zöngés hangok esetén kvázi-periodikus, zöngétlen hangok esetén zajszerű. A gerjesztés alakja a hangszínezetet befolyásolja. Egy formáns szűrő a megadott formáns frekvencia környezetében erősíti a jelet, ezzel modellezve a garat, a gége és a szájüreg által alkotott rezonátor-rendszer erősítéseit. A formáns frekvenciák a zöngés hangokra jellemzőek, de zöngétlen hangok leírására is használhatóak.

Az első három formánsfrekvencia jól leír egy zöngés, kitartott beszédhangot. A szintézis folyamán beállítandó formánsfrekvenciákat a szövegből előállított fonémasorozat határozza meg. A formánsfrekvenciák az artikulációs mozgások függvényében szintén változnak, a spektrális szerkezet folymatosan módosul periódusról periódusra.

A beszédhangokon belül megkülönböztetünk stabil szakaszt (általában a hang közepe) és hangátmeneti részt (a hangnak a kezdete, ami az előző hanghoz kapcsolódik, illetve a vége, amelyik a következőhöz fűződik hozzá). A stabil szakaszok közötti hangátmeneteknél a formánsok mozgatását a bemeneti fonémasorozat és hangidőtartamok alapján szabályrendszer vezérli. A szabályok komplexitási szintje meghatározza a szintetizátor hangzását. Egyszerű szabályokkal csak gépies hangzás érhető el. Az újabb formánsszintetizátorokban a hangátmenetek paramétereit természetes hangátmenetekből nyerik ki. Ez valamivel jobb hangzást eredményez.

1. ábra
Soros elrendezésű formánsszintetizátor blokkvázlata



A zöngés gerjesztés alapfrekvenciájának a vezérlését az úgynevezett prozódiai modul végzi. Ennek bemenete a szövegből előállított fonémasorozat és a szimbolikus prozódia. Ez utóbbi általában a mondatok modalitását (kijelentő/kérdő) és a hangsúlyok mondaton belüli helyét és típusát adja meg. A modul kimenetei a fizikai prozódiai jellemzők, azaz az alapfrekvencia-menet, a fonémáknak megfelelő hangok időtartamai, illetve az intenzitásmenet. Ezek minősége szintén nagyban befolyásolja az előállított beszéd hangzásának természetességét.

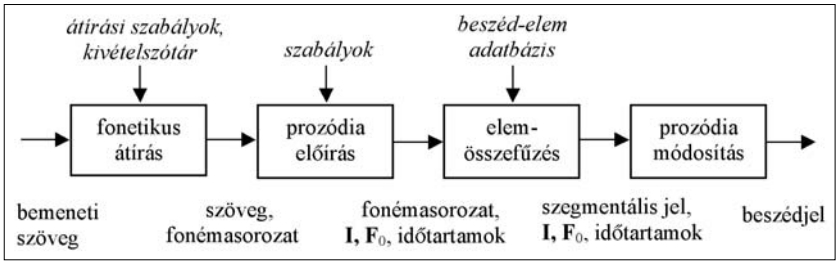
A formánsszintetizátor megfelelő vezérlésével jó minőségű, természetes beszéd állítható elő. Ugyanakkor ilyen vezérlő információt csak természetes beszédjelből, félautomatikus módszerek segítségével sikerült mindezidáig előállítani. A bemeneti szövegből kiinduló és egy szabályhalmaz segítségével előállított vezérlő információ érthető, de erősen gépies hangzású beszéd előállítását teszi csak lehetővé. Ezen minőségi korlát miatt a formánsszintetizátorok csak kis erőforrásigényű gyakorlati alkalmazásokban fordulnak elő. A módszert kutatási célra ma is használják, elsősorban azért, mert a beszédjel gerjesztése – ellentétben az újabb beszédszintézis-technológiákkal – könnyen módosítható, és így annak hatása külön vizsgálható. Másik előnye a kis tárkapacitás-, és az alacsony számításigény.

A BME TMIT által kifejlesztett Multivox 12 nyelven beszélő formánsszintetizátor magyar nyelvű változata [1] ingyenesen hozzáférhető.

3. Elemösszefűzésen alapuló beszédszintézis

Az elemösszefűzésen alapuló beszédszintézis esetében a szövegfelolvasó rendszer két fő egysége közül (szövegfeldolgozó és hullámforma-előállító) a hullámforma-előállító rész jelent újdonságot *(2. ábra)*. Ahelyett, hogy minden egyes beszédhangra és beszédhangátmenetre előírnánk a formánsfrekvenciák és a gerjesztés alakjának időbeni menetét, természetes beszédből kivágott hullámforma elemeket fűzünk össze. Ugyanakkor a formánsszintetizátorban alkalmazott prozódiai modul általában változatlanul megmarad, és előírja az előállítandó hullámforma alapfrekvencia- és intenzitásmenetét, illetve az egyes hangok időtartamait. Emellett a szöveget fonémasorozattá, illetve szimbolikus prozódiává alakító rész is változatlan marad.

2. ábra Elemösszefűzésen alapuló beszédszintetizátor egy lehetséges felépítése



A technológia egyik alapvető kérdése, hogy melyek legyenek azok a hullámforma elemek, amelyek összefűzésével előáll a gépi beszéd. Itt több szempontot kell figyelembe venni. Egyrészt teljes fedést kell biztosítani, azaz az adott nyelv tetszőleges hangsorozatát elő kell tudni állítani. Másrészt az előállított beszédnek minél természetesebben kell szólnia. Korlátot jelenthet a hullámforma elemek száma, illetve azok együttes mérete. Az előbbi az elemek közötti keresés idejét növeli, az utóbbi pedig a szükséges tárterületet.

Alapötletként felmerülhet a fonémáknak megfelelő hangok, mint elemek használata. Ez teljes fedést biztosít, és kevés elemmel megoldható (a magyarra 38 fonémából már előállítható jó minőségű beszédszintetizátor). A fonémáknak megfelelő hangok összefűzésével előálló jel azonban nem hangzik folytonosnak, azaz az összefűzés után előálló hang minősége gyenge. A problémát az okozza, hogy a beszédjelben az egyes hangok folyamatosan mennek át egymásba, és általában csak a hang közepén tekinthetők állandósultnak. Ezt úgy is értelmezhetjük, hogy a hang elejének alakulása az őt megelőző, a vége pedig az őt követő hangtól függ [2]. A megoldás a környezetfüggő hangok használata lehetne, ahol minden egyes hangot minden lehetséges hangkörnyezetének megfelelő változatban tárolunk. Ehhez viszont nagyon sok elemet kellene tárolni, felső becslésként 38^3=54872 elemre lenne szükség. Ez azért felső becslés, mert a nyelvben nem valósul meg minden lehetséges hanghármas. Az ilyen környezetfüggő elemeket *triád*nak (angolul triphone-nak) nevezik. Megjegyezzük, hogy az elemszámot nemcsak a technológia korlátozza, hanem a rendelkezésre álló emberi erőforrások is. Az egyes elemeket ugyanis egyenként kell hangstúdióban felvenni, és félautomatikus módszerekkel feldolgozni, előkészíteni a szintézisre.

A gyakorlatban bevált kompromisszumos megoldás a két egymás utáni félhang együtteseként előálló diád (angolul diphone) alkalmazása. Ez kihasználja azt a közelítő feltételezést, hogy egy hang első fele nem függ az azt követő hangtól, második fele pedig az azt megelőzőtől. A magyar nyelvű szintézishez szükséges diád-elemek száma 38^2=1444. Megjegyezzük, hogy a kezdeti, csak diád-elemeket tartalmazó rendszereket az idők folyamán triád-elemekkel bővítették, ami némi minőség javulást eredményezett. Ezek a triád-elemek az adott hangot megelőző hang közepén kezdődtek és a hangot követő hang közepéig tartanak, azaz két hangnyi hosszúak. Ennek előnye, hogy a rendszer kevesebb vágási pontot tartalmaz, a hosszabb építőelemek miatt pedig folytonosabb lesz a beszéd hangzása. Hátránya a megnövekedett elemszám.

Hasonló utat járt be a BME TMIT-en kifejlesztett Profivox magyar nyelvű beszédszintetizátor [3], amelynek legújabb változata az 1444 diád mellett, 6000 triád-elemet is tartalmaz. A két rendszer hangzását összehasonlító egyszerű minősítési teszt megtalálható [4]-ben.

A diád, illetve triád-elemek összefűzése után gondoskodni kell arról, hogy az előálló beszédjel kövesse a formánsszintézisnél alkalmazott módhoz hasonlóan előírt prozódiai jellemzőket (alapfrekvencia- és intenzitásmenet, hangidőtartamok). A formánsszintézissel ellentétben itt nem áll eleve rendelkezésre egy parametrikus modell, így azt vagy létre kell hozni, vagy valamilyen időtartománybeli manipuláció segítségével kell a jelet módosítani. Mindkét esetben szükséges a jel alapfrekvencia-mentének pontos ismerete. Az egyes elemek esetleges modell-paraméterei, illetve alapfrekvencia-menete előre, nem valós időben is meghatározhatóak. Általánosan igaz, hogy minél nagyobb mértékben módosul a beszédjel, annál jobban romlik a minősége. Az intenzitás-menet viszonylag szabadon módosítható, ugyanakkor ennek van a legkisebb szerepe a prozódia alakításában. A hangidőtartamok akár 50-200%-os tartományban módosíthatóak, ami a gyakorlatban elegendő. Az előírt alapfrekvencia-menet megvalósítása a legkritikusabb, ugyanis az alapfrekvencia csak körülbelül 30%-kal módosítható még elfogadható minőségben.

A technológiához hozzátartozik a hangfelvételek rögzítése és feldolgozása. Az egyes diád-, illetve triád-elemeket mesterséges szavakba, úgynevezett logatomokba (például „aboka”: a „b-o” diád-elemhez) ágyazva kell felolvasni. Ezeket a bemondónak jól artikulálva, monoton hanglejtéssel kell felolvasnia. A logatomok biztosítják, hogy minél kevésbé érvényesüljön a szomszédos hangok hatása a diád-elemre [5]. Ugyanakkor ezeket összefűzve mellékhatásként „túlartikulált” lesz a beszéd hangzása. A monoton hanglejtésre azért van szükség, hogy az alapfrekvencia módosításkor ne kelljen az esetlegesen túl alacsony, vagy túl magas frekvenciájú elemeken sokat módosítani. A felvett logatomok hullámformájában félautomatikus módszerekkel kell a hanghatárokat jelölni, és az egyes diád-, illetve triád-elemeket kivágni, ellenőrizni és hibás ejtés esetén javítani. Egy diád-elemeket tartalmazó hangadatbázis felvétele körülbelül 4 órát, míg feldolgozása egy emberhónapot vesz igénybe.

A diád-, illetve triád-elem összefűzésen alapuló beszédszintézis technológiát elterjedten alkalmazzák. Erre példa a Profivox magyar nyelvű beszédszintetizátor [3], amely e-mail-felolvasó, SMS-felolvasó, számszerinti tudakozó, illetve egyéb alkalmazásokban működik. A csak diád-elemeket tartalmazó változat kis memóriaigényének köszönhetően Symbian- és Windows Mobile-alapú mobiltelefonokon is képes futni.

4. Korpusz alapú, elemkiválasztásos beszédszintézis

Az elemösszefűzéses technológia továbbfejlesztéseként jött létre a korpusz alapú, elemkiválasztásos beszédszintézis *(3. ábra)*.

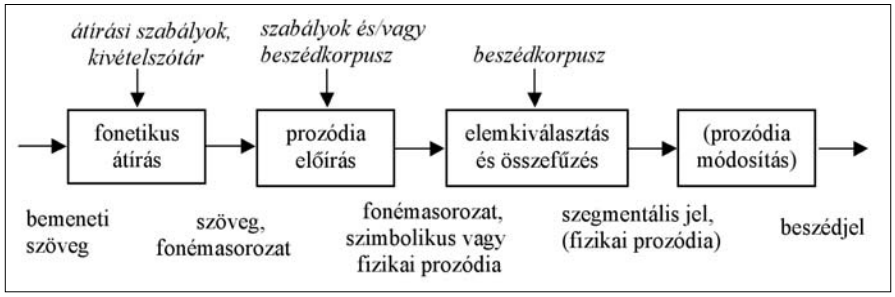
A technológia továbbra is beszédelemek összefűzésen alapul, ugyanakkor – mint a neve is mutatja – további két elvet vezet be. Az első, a korpusz alapú szintézis elve szerint a beszédszintetizátor hangadatbázisa nem a monoton prozódiájú logatomokból kivágott diád–, illetve triád-elemeket, hanem természetes hangzású teljes mondatokat tartalmaz. A mondatok egy nagyméretű szövegkorpuszból kerülnek kiválogatásra, és azok felolvasásával jön létre a több órányi beszédet tartalmazó adatbázis, azaz a beszédkorpusz.

Ellentétben a hagyományos elemösszefűzésen alapuló technológiával, a korpuszos adatbázisban egy adott hangsort tartalmazó beszédelem általában több példányban is előfordul. Ezen példányok prozódiai megvalósítása (alapfrekvencia-, és intenzitás-menet, hangidőtartamok, hangszínezet) eltérő. Másrészt a beszédkorpuszban egyszerre több különböző méretű elem is definiálható (például diád, triád, szótag, szó stb.). Ezen két ok miatt több lehetséges módon állítható elő egy adott szintetizált beszédszakasz, amelyek közül a legtermészetesebben hangzó változatot kell kiválasztani. Ezt a folyamatot elemkiválasztásnak nevezik, arra utalva, hogy a többféle lehetséges elem közül kiválasztjuk, hogy melyek kerüljenek összefűzésre egy adott bemondás előállításához. Megvalósítása a hibajavító kódolásban és a beszédfelismerésben is alkalmazott Viterbi algoritmus segítségével történik.

Megjegyezzük, hogy a hagyományos elemösszefűzésen alapuló beszédszintézis rendszernél is szükség van elemkiválasztásra, ha az adatbázis vegyesen tartalmaz diád- és triád-elemeket. Ugyanakkor – mivel minden elem csak egy változatban, egy adott prozódiával szerepel – a kiválasztás kevesebb számítással megoldható [6].

A korpusz alapú elemkiválasztásos beszédszintézis alapvetően két ok miatt eredményez jelentős minőségjavulást a hagyományos elemösszefűzéshez képest. Egyrészt kevesebb összefűzési pontot tartalmaz, mint a logatomos diád-, triád-elemekből építkező rendszer, ami folytonosabb, természetesebb hangzást eredményez [2]. Másrészt a megfelelő prozódia kialakításához kevesebb jelfeldolgozási művelet szükséges, mivel az adatbázis elemek prozódiai változatossága miatt általában ki tudunk olyan elemeket választani, amelyek közel állnak a kívánt prozódiához. Emellett a diád-elemeknél jóval hosszabb, egybefüggő beszédda-

3. ábra
Korpusz alapú, elemkiválasztásos beszédszintetizátor felépítése



rabok természetes prozódiája is megőrizhető, illetve (ha egyáltalán szükséges a prozódia módosítása) viszonylag egyszerűen eltolással biztosítható az egybefűzött darabok prozódiai illeszkedése. Ez természetesebb hangzást eredményez, mint a szabály alapon előírt, a diád-elemekből összerakott jelre kényszerített mesterséges prozódia.

A korpusz alapú szintézis velejárója a beszédadatbázis méretének jelentős növekedése. Az *1. táblázat* mutatja az adatbázis méretének változását az egyre jobb minőséget biztosító beszédszintézis technológiák esetén. A nagyobb tárigény mellett az elemkiválasztás nagyobb számításigényt is támaszt a korábbi rendszerekhez képest. Ezeket részben kompenzálja a számítástechnikai eszközök (háttértár, CPU) időközben végbement fejlődése.

A diád-, illetve triád-elemek összefűzésén alapuló szintézishez hasonlóan, itt is szükséges az elemhatárok jelölése a felvett beszédadatbázisban. Az 1. táblázatból látható, hogy az adatbázis időtartama, és ezzel arányosan a benne szereplő hangok száma több mint egy nagyságrenddel megnőtt. Az elemhatárok jelölése már nem végezhető a korábbi, sok kézi munkát igénylő módszerrel. Szerencsére a gépi beszédfelismerés technológiájának időközben bekövetkezett fejlődése lehetővé teszi az elemhatár-jelölés automatizálását.

A korábbi, formáns szintézisen, illetve a diád- és triád-elemek összefűzésén alapuló rendszerek minősége nem függött a felolvasandó szöveg tartalmától. A korpusz alapú szintézis minőségét nagyban befolyásolja, hogy a szintetizálandó szöveg mennyire van közel a szintetizátor beszédkorpuszának, azaz adatbázisának tartalmához. Minél közelebbi a szintetizálandó szöveg, annál nagyobb beszéd-darabokat lehet egyben kivenni az adatbázisból, megőrizve azok természetességét. Adott esetben előfordulhat az is, hogy a szintetizálandó szöveg egy-egy mondatát teljes egészében megtaláljuk. Ellenpéldaként viszont lehet, hogy különálló diád- vagy triád-méretű elemekből kell összerakni egy, az adatbázisban nem szereplő szövegrészt, megnövelve a vágási pontok számát. A fenti ok miatt célszerű a szintetizátor beszédkorpuszát a felolvasandó szövegekhez, azaz az adott alkalmazáshoz igazítani (például csak időjárási szövegekből álló korpusz időjárásjelentés felolvastatásához).

5. A BME TMIT kísérleti, korpusz alapú beszédszintetizátora

A BME TMIT-en elkészült egy magyar nyelvű korpusz alapú elemkiválasztásos beszédszintetizátor kísérleti változata. Korábban már beszámoltunk a fejlesztés kezdeti lépéseiről [4]. Itt az azóta elvégzett munka eredményeit ismertetjük.

Célunk első lépésben nem egy általános beszédszintetizátor kifejlesztése, hanem egy specifikus alkalmazás, egy időjárás-jelentés felolvasó megvalósítása volt. A későbbiekben a fejlesztés során szerzett tapasztalatainkat használjuk majd fel egy általános célú, korpusz alapú elemkiválasztásos beszédszintetizátor megvalósításához.

A kísérleti rendszer beszédkorpuszát Internetről gyűjtött időjárás-jelentés szövegekből állítottuk össze (ennek részleteit lásd [2]-ben). A szöveganyag 5400 mondatát egy fiatal színésznő olvasta fel, amit stúdió körülmények között, 44,1 kHz-es mintavételi frekvenciával, mintánként 16 biten rögzítettünk. A felvételek 4 héten keresztül zajlottak, heti 2-3 nap terjedelemben. Az 5400 mondat felvételével mintegy 11 órányi folyamatos beszédből álló hanganyag állt össze.

5.1. Hangfelvételek címkézése

Az adatbázis hanganyagát mondatokra bontottuk szét. Minden egyes mondathoz tartozik egy szöveges átírás. Ezt a mondat felolvasásához használt szövegből származtattuk. A mondatokban az elemhatárok, illetve a zöngeperiódus-határok jelölése automatikusan

Szintetizátor technológia	Adatbázis tartalma	Mintavételi frekvencia	Adatbázis mérete	Hangfelvétellel előálló adatbázis időtartama	Hangfelvétel időigénye
Multivox formáns szintetizátor *(1985)*	12 paraméterrel meghatározott hangszeletek (4 formáns)	11,025 kHz	1,2 K	parametrikus szintézis, nincs hangfelvétel	parametrikus szintézis, nincs hangfelvétel
Profivox diád-elemes szintetizátor *(1997)*	1444 diád-elem	8 kHz/ 22,05 kHz	1,5 MB/ 7 MB	3 perc	4 óra
Profivox diád- és triád-elemes szintetizátor *(2002)*	1444 diád- és 6000 triád-elem	22,05 kHz	92 MB	35 perc	10 óra
Kísérleti korpusz alapú szintetizátor *(2006)*	5400 mondat	22,05 kHz	1,7 GB	11 óra	50 óra

1. táblázat

A BME-TMIT-en fejlesztett beszédszintetizátorok adatbázisméretének növekedése

történt. A zöngeperiódusok a beszéd zöngés (kvázi-periodikus) részének egy-egy periódusát jelentik. A zöngétlen részeken a beszéd nem periodikus, ezért ott 5 ms-onkénti jelöléseket alkalmaztunk.

A zöngeperiódus-határok jelölése egyrészt az alapfrekvenvcia pillanatnyi értékének a meghatározásához kell, másrészt az alapfrekvencia-menet esetleges módosításához szükséges. A jel alapfrekvencia-menete kiszámítható a periódusidők reciprokaként. Az alapfrekvencia-menetet az elemkiválasztás folyamán is felhasználjuk. A zöngeperiódus-határok bejelöléséhez a Praat fonetikai és beszéd-analizátor szoftverben implementált ablakfüggvénnyel korrigált autokorrelációs módszeren alapuló alapfrekvencia-detektálást használtuk fel [7].

A szintetizátorban kétféle építőelemtípust definiáltunk, ezek a szó és a beszédhang. A beszédhang biztosítja a teljes fedést, azaz, hogy tetszőleges tartalmú szöveget elő lehessen állítani. A szó szint gyorsabb keresést tesz lehetővé az adatbázisban és biztosítja a bemondó hangszínetéhez közel álló hangzás elérését. Ennek megfelelően a felvett beszédkorpuszban jelölni kell a hang és szóhatárokat. A hanghatárok jelölése a beszédjel szintjén történik, azaz megadjuk, hogy hányadik mintán kezdődnek az egyes hangok. Ezt a feladatot egy a BME TMIT-en kifejlesztett magyar nyelvű beszédfelismerő [8] segítségével oldjuk meg. A felismerőt kényszerített módban használjuk, ami azt jelenti, hogy a beszédet tartalmazó hangfájl mellett bemenetként megadjuk annak szöveges tartalmát is, ami meghatározza, hogy milyen hangsorozatokat keressen a felismerő.

Ehhez először a szöveg fonetikus átírását kell elvégezni, ami automatikusan történik, a magyar nyelv hasonulási szabályainak figyelembe vételével. Megjegyezzük, hogy a szövegben szereplő rövidítések, illetve az esetleges speciális jelek (pl. mínusz jel) átírását kézzel kell elvégezni és ellenőrizni. A leírt szöveg nem definiálja egyértelműen a megvalósuló hangsorozatot. Egyrészt előfordulhatnak kiejtési variációk (de ezek inkább csak a spontán beszédre jellemzőek), másrészt a szóhatároknál csak opcionálisan tartunk szünetet. Az is előfordul, hogy szóhatáron átívelő hasonulás, vagy egybeolvadás jön létre. Ezekre példa a „hűvös záporok”, ami legtöbbször „hűvözs záporokként” kerül kimondásra, illetve az „ideig ködös” szókapcsolat, ahol általában egy darab hosszú „k” hang valósul meg a szóhatáron. Ennek kezelésére a felismerő a fonetikus átírás többféle változatát állítja elő, amit egy irányított gráf ír le.

Jelen változatban a gráf csak a szóhatároknál ágazik ketté egy, az adott szóhatáron szünetet tartalmazó, illetve egy szünetet nem tartalmazó hangsorozattá. A szünetet nem tartalmazó változatban működnek a hasonulási és egybeolvadási szabályok. Ezek az alternatív útvonalak itt nem részletezett módon beépülnek a rejtett Markov-modelleket használó beszédfelismerő keresési terébe, és ezek közül a beszédjel alapján a legvalószínűbb hangsorozat kerül kiválasztásra a Viterbi algoritmus felhasználásával. A kényszerített felismerés nem csak a hanghatárokat, hanem a beszédjelben megvalósult hangsorozatot is megadja kimenetként.

A felismerő 20 ms-os keretekkel és 10 ms-os kereteltolással dolgozik, azaz a hanghatárokat elvileg is csak 10 ms-os pontossággal határozza meg. A hanghatárokat (zöngés hangok esetén) a legközelebbi zöngeperiódushoz igazítjuk, ezáltal biztosítjuk, hogy szintetizált mondatban egymás mellé kerülő beszéd-darabok azonos fázisban legyenek, azaz teljes periódusokból álljanak. Így nem lesz pattogó, vagy rekedtes hangzású az előállított beszéd.

A hanghatár-jelölés ellenőrzéséhez minden egyes hangra egy hanghossz eloszlás hisztogrammot készítettünk. Ennek segítségével meghatároztuk azokat a hangokat, amelyek hossza jelentősen eltért a velük azonos hangok átlagolt hosszaitól. Az ilyen hangokat tartalmazó mondatokat külön-külön manuálisan megvizsgáltuk. A tapasztalt hibákból sorolunk fel néhányat.

Az abnormális hanghosszak egy része átírási hibákból származott, azaz a hangfájlok és a hozzájuk tartozó szöveges fájlok tartalma nem minden esetben felelt meg egymásnak. Az is előfordult, hogy a szövegben olyan rövidítések maradtak, amelyeket nem tudott megfelelően feloldani a fonetikus átíró.

A hibák másik részét az automatikus hanghatár-jelölés okozta. Tipikus hiba például a szóvégi réshangok (“f”, „sz”) és a szavak közötti esetleges levegővételek egymásba csúszása. Hasonlóan a „c” hang határa is sok esetben a szomszédos hangra csúszott. Ennek valószínű oka, hogy a beszédfelismerő tanítása egy telefonon keresztül felvett adatbázissal történt, amelyben az átviteli tulajdonságok miatt a „c” hang spektrumának nagy része elveszett. Ennek megoldására a felismerőt magával a felcímkézendő hangadatbázissal kell betanítani, ami várhatóan pontosabb hanghatárokat fog eredményezni.

A korábbi tapasztaltok alapján a 11 órányi folytonos beszédkorpusz elegendő hosszú a felismerő megfelelő tanításához.

A szóhatárok jelölését a beszédfelismerő által visszaadott, a beszédjelhez legjobban illeszkedő fonémasorozaton végeztük. A szavakon átívelő hangegybeolvadások (például: „ideig ködös”) miatt előfordulhat, hogy egy hang egyszerre két szóhoz is tartozik. Ennek kezelésére külön jelölést alkalmaztunk a szavak kezdetére és végére, így lehetővé tettük a szavak közötti átfedést (például: „<idei<k>ödös>”, ahol „<” a szókezdetét, „>” a szó végét jelöli).

A szóhatárok jelölését teljesen automatikusan végeztük, olymódon, hogy a beszédfelismerő által visszaadott fonémasorozatot illesztettük a szintén a beszédfelismerő által előállított, (a szóhatároknál elágazó) összes lehetséges fonetikus átírást megadó irányított gráfhoz. Az illesztést egy állapotgéppel végeztük, amely a fonémasorozat és a gráf alapján követte, hogy mikor merre ágazott el a felismerő, és az elágazásoknál a választásnak megfelelően szúrta be a szóhatárokat.

5.2. Prozódia

A kísérleti korpusz alapú szintetizátor első lépésben a szöveg fonetikus átírását végzi, ami nem változott a korábbi szintetizátorokhoz képest. A fonetikus átírást követően a prozódia előírására kerül sor. Az alábbiakban ennek lehetséges megvalósítási módszereit elemezzük.

Az elemösszefűzésen alapuló beszédszintetizátor szabályalapú prozódiája alkalmazható a korpusz alapú, elemkiválasztásos rendszerben is. A megközelítés előnye, hogy a szabályok átvehetők a korábbi rendszerből, hátránya viszont, hogy a megvalósított prozódia ugyan elfogadható, de nem ad természetes hangzást. A szabályalapon meghatározott prozódia először szimbolikus, majd fizikai szinten áll elő. A korábbi, diádos, triádos elemösszefűzésen alapuló technológiánál a monoton elemek egymásután illesztésével előálló beszédjelre kellett a fizikai-szintű prozódiát ráültetni. A korpusz alapú rendszernél azonban lehetőség nyílik arra, hogy az előírt prozódiát a jel összefűzése előtt, az elemkiválasztás folyamán vegyük figyelembe. Ez történhet fizikai szinten, de akár a szimbolikus prozódia szintjén is. Az előbbi azt jelenti, hogy például az előírt alapfrekvencia-menethez minél közelebbi elemeket választunk ki. Az utóbbi pedig azt jelenti például, hogy ha a bemeneti szöveg alapján előírt szimbolikus prozódia szerint hangsúlyos egy szó, akkor a beszédkorpuszban olyan elemet, vagy elemeket keresünk, amely az előírás szerinti hangsúllyal rendelkezik. Ennek fizikai megvalósítása számos problémát vet fel, amiket itt nem részletezünk.

A szimbolikus prozódia alapján történő elemkiválasztás előnye, hogy megőrződik a kiválasztott elemek természetessége, hátránya, hogy a vágási pontoknál megtörhet a természetes prozódia, ezért utólagos prozódia-simításra van szükség. Ez azt jelenti, hogy az előállított mondatban szereplő beszédelemek alapfrekvencia-, és intenzitás-menetét, illetve hangidőtartamait jelfeldolgozási módszerekkel úgy módosítjuk, hogy az egymás melletti elemek között folytonos legyen az átmenet. Megjegyezzük, hogy abban az esetben is szükség lehet ilyen simításra, ha a fizikai szintű prozódiát használjuk az elemkiválasztás folyamán. Továbbá ekkor sem feltétlenül kell jelfeldolgozással pontosan a jelre kényszerítenünk az előírt fizikai prozódiát.

A szimbolikus prozódia használata esetében probléma, hogy mind a szintetizátor beszédkorpuszában, mind a szintetizálandó szövegben jelölni kell a szimbolikus információt (például a hangsúlyokat). Ezt a jelenlegi automatikus módszerek csak pontatlanul tudják megtenni. A kézi jelölés a beszédkorpusz mérete miatt nem praktikus megoldás. Ráadásul nem garantált a konzisztencia a korpuszban kézileg jelölt hangsúlyok, illetve a bemeneti, szintetizálandó szövegből előre jelzett hangsúlyok között. A szimbolikus szinthez képest még egy szinttel magasabb információkat is használhatunk az elemkiválasztás során. A szimbolikus prozódiát a bemeneti szöveg (leegyszerűsített felszíni) nyelvi elemzése alapján határozzuk meg. Ennek a nyelvi elemzésnek a kimenete alapján is kereshetünk a beszédkorpuszban. A módszer használata esetén szintén szükséges az összefűzött elemek prozódiai simítása.

További lehetőség a korpusz alapú fizikai prozódia generálása. A korpusz mondatainak fizikai prozódiáját ki lehet nyerni a korpuszból. Ez azt jelenti, hogy a szimbolikus prozódiából nem csak szabályok segítségével tudjuk előállítani a fizikai prozódiát, hanem a szimbolikus prozódia alapján a beszédkorpuszban, mint fizikai prozódia tárban is kereshetünk. A keresés történhet a beszédkorpusz mondataihoz tartozó (előre meghatározott), illetve a szintetizálandó szöveghez meghatározott szimbolikus prozódia egyezése alapján. Természetesen itt is elképzelhető a nyelvi elemzés szintjére történő visszalépés, azaz a nyelvi elemzés által megadott információk alapján történő keresés. Az adatbázisból a keresés során kinyert fizikai prozódiát előírhatjuk az előállítandó mondat cél-prozódiájaként.

A kísérleti rendszerben leegyszerűsített szimbolikus prozódiát használunk, szópozíció jellegű információ formájában. Az elemkiválasztás előtt a bemeneti szöveget prozódiai egységek szerint tagoljuk. A prozódiai egység a szintetizátor jelenlegi megvalósításában egy írásjelekkel határolt, tagmondat-jellegű szövegrészt jelent. Minden egyes prozódiai egységet megcímkézünk aszerint, hogy a mondaton belül milyen pozícióban van (első-, utolsó-, közbenső szó).

A szópozíció százalékos formában megadja az adott szó helyét az azt tartalmazó prozódia egységben. Ezt a prozódiai egység fonémákban megadott hossza és a szó első/utolsó fonémájának prozódiai egységen belüli pozíciója alapján megállapított százalékos értékek definiálják. Az elemkiválasztás folyamán megpróbálunk a bemeneti szövegben szereplő szavakhoz hasonló pozíciójú szavakat kiválasztani. Természetesen a pozíció jellegű információ nem határozza meg egyértelműen sem a hangsúlyokat, sem a hangidőtartamokat. A módszer előnye az egyszerűsége, hátránya, hogy sok esetben nem biztosít megfelelő prozódiát. A kísérleti rendszer jelenlegi implementációja nem tartalmaz utólagos prozódia simítást.

5.3. Elemkiválasztás és összefűzés

A prozódia előírása után, a kísérleti korpusz alapú szintézis rendszer harmadik lépése az elemkiválasztás. Az elemkiválasztás alapelve, hogy a rendszer a bemeneti szöveget (elviekben) az összes lehetséges módon összerakja a beszédkorpusz elemeiből, és azok közül a legtermészetesebben hangzót választja ki. A természetesség automatikus megállapításához kétféle költséget vezetünk be.

Az egyezési költség megadja, hogy egy adott elem mennyire felel meg a szintetizálandó beszédszakasznak. A jelenlegi megvalósításban a beszédszakaszt annak betűsorozatként megadott szöveges tartalma, illetve hangsorozatként megadott fonetikus átírása határozza meg. Ehhez járulnak még a szintetizálandó szövegből meghatározott prozódiai előírások, ami jelenleg a szó-, illetve hangpozíciókat jelenti. Az összefűzési

költség azt adja meg, hogy a leendő szomszédos elemek mennyire folytonosan illeszkednének egymáshoz. Az elemkiválasztás folyamata azt a mondatot választja ki az összes lehetséges közül, amelyre az egyezési és összefűzési költségek összege a legkisebb.

A kísérleti szintetizátor kétféle elemet, szavakat és beszédhangokat kezel. Szószintű keresés esetén az adott szót alkotó betűsorozat alapján azonosítjuk az elemeket, míg beszédhangszintű keresés esetén a hangot megadó fonéma alapján. Az elemkiválasztás hierarchikusan történik. Első menetben csak a szószintű elemek között keres a rendszer. Ha a bemeneti szövegnek vannak olyan szavai, amit nem sikerült szóalapon lefedni, akkor a hiányzó szavakat beszédhangokból rakja össze a rendszer. A szószinten már megtalált elemeket nem próbáljuk kisebb elemekből előállítani, még akkor sem ha ez esetleg prozódiai szempontból célszerű lenne.

A szószint bevezetésének alapvető előnye a gyorsabb keresés. Ennek hatékonysága függ a szintetizátor beszédkorpusza és az előállítandó szöveg közötti hasonlóságtól. Egy általános célú beszédszintetizátornál ez kevésbé hatékony megoldás – különösen a ragozó magyar nyelv esetében – ugyanakkor korlátozott tematikájú alkalmazások, például időjárásjelentés-felolvasás esetén jól működik. Általános célú alkalmazásnál is gyorsítható a keresés a beszédhangnál hosszabb elemek, például szótagok bevezetésével.

Az elemkiválasztást mondatonként végezzük. A keresés folyamán az adott mondatban szereplő egy-egy szóhoz, vagy hanghoz többféle lehetséges jelöltet is kiválasztunk. Egy-egy elemhez implementációs és hatékonysági okokból maximáltuk a lehetséges jelöltek számát. Ha a kiválasztás folyamán egy adott elemhez tartozó jelöltek száma elérte a megadott maximumot, akkor minden egyes további jelölt hozzávétele után a legmagasabb egyezési költségű elemet eldobjuk. Az összefűzés során elvileg minden elem esetében tetszőleges jelöltet kiválaszthatunk. Ebből következőleg a különböző előállítható lehetséges mondatok számát a jelöltek számának szorzata adja meg. Az optimális mondat kiválasztását a dinamikus programozáson alapuló Viterbi-algoritmus segítségével végezzük. Az algoritmus minden egyes lehetséges útra (mondatra) meghatározza az egyezési és összefűzési költségek összegét, és a minimális költségű utat választja ki.

Az egyezési költségben az alább felsoroltak szerepelnek:

1. A jelöltet (szó vagy beszédhang) megelőző és követő fonéma egyezése az előírt célelemet megelőző és követő fonémával. A legkisebb költséget a teljes egyezés jelenti. Emellett definiáltunk egymással helyettesíthető fonéma-kategóriákat [2]. Az azonos kategóriába eső hangok egyezési költsége kisebb, mint az eltérő kategóriákba esőké. Eltérő kategóriák esetén az egyezési költséget egy költségmátrix definiálja, amelynek értékei minden esetben nagyobbak, mint azonos kategória esetén.
2. Prozódiai egység mondaton belüli pozíciójának egyezése. Ezt csak szavak esetében vesszük figyelembe.
3. Prozódiai egységen belül előírt pozíciótól való eltérés. Ezt csak szavak esetében vesszük figyelembe.

Az összefűzési költség a következők szerint alakul:

1. Ha a vizsgált két jelölt a beszédkorpuszban egymás után következett, akkor az összefűzési költség mindig 0.
2. Ha a vizsgált két jelölt a beszédkorpuszban azonos mondatból származott, akkor kisebb összefűzési költséget rendelünk hozzá, mintha eltérő mondatokból származott volna.
3. Alapfrekvencia-menet folytonossági költsége, amit az első elem végső és a második elem kezdő alapfrekvenciájának eltérésével arányosan számolunk.

Az egyes költségek értékeit számos hangminta meghallgatása során, ad hoc módon állítottuk be. Ezek optimalizálása még tovább javíthatja az előállított beszéd minőségét.

6. Szintézis rendszerek három generációjának összehasonlítása szubjektív minősítéssel

A következőkben leírt vizsgálatunk célja annak a meghatározása volt, hogy mekkora a minőségi ugrás a különböző generációjú beszédszintézis rendszerek között. Emellett alaposabban akartuk vizsgálni a kísérleti korpusz alapú, elem-összefűzéses szintetizátorral előállított mondatok minőségét.

Korábban már ismertettünk egy hasonló tesztet a diád- és triád-elemek összefűzésén alapuló technológia, és a korpuszos, elemkiválasztáson alapuló szintetizátor összehasonlítására [4]. Az ott ismertetett tesztben a korpuszos rendszer működését kézzel összevágott mondatok szimulálták, mivel akkor még nem állt rendelkezésre működő kísérleti rendszer.

A beszédszintézis rendszerek minőségét meghallgatásos tesztek során végzett szubjektív minősítéssel lehet összehasonlítani. Ennek egyik módja a MOS (Mean Opinion Score – átlagos szubjektív osztályzat) teszt alkalmazása. A teszt során a tesztelők véletlenszerű sorrendben hallgatják meg a különböző szintetizátorokból származó mondatokat, és azok minőségét egyenként osztályozzák egy ötfokozatú skálán *(2. táblázat)*. Az osztályzatok összes tesztelőre vonatkoztatott átlaga adja meg a MOS értékét.

Megjegyezzük; a beszédszintetizátoroknál hagyományosan az érthetőséget szokás vizsgálni, ez azonban az újabb rendszereknél kevésbé okoz problémát.

5	kiváló
4	jó
3	közepes
2	gyenge
1	rossz

2. táblázat

Ötfokozatú szubjektív hangminősítő skála

Továbbá létezik a beszédszintetizátorok minősítésére vonatkozó P.85-ös ITU-T szabvány [9], amely más szempontokat, például a szintetizált szöveg megértéshez szükséges koncentráció mértékét is vizsgálja. Ez a gyakorlatban nem terjedt el. Ennek lehetséges magyarázatát kereste egy tanulmány [10], amely kimutatta, hogy a különböző vizsgált szempontokra vonatkozó minősítések nagy korrelációt mutatnak, azaz valójában nem adnak plusz információt az egyszerű minősítéshez képest, viszont fölöslegesen növelik a teszt idejét és költségét.

A jelen tesztben három, a BME TMIT-en fejlesztett magyar nyelvű beszédszintetizátort hasonlítottunk össze. Az első a Multivox formánsszintetizátor női hangú változata, a második a Profivox szintetizátor diád- és triád-elemeket tartalmazó változata volt. Ennek adatbázisa ugyanattól a női bemondótól származott, mint a vizsgálatban szereplő harmadik, kísérleti korpusz-alapú, elem-összefűzéses beszédszintetizátoré. (Megjegyezzük, hogy a Multivox és a Profivox szintetizátor – tapasztalataink szerint – férfi-hangú adatbázissal némileg jobb minőségű beszédet ad, mint a női adatbázisokkal. Ez megegyezik a nemzetközi tapasztalatokkal, és a hangok eltérő fizikai jellegzetességeiből adódik. Például a magasabb frekvenciájú hang minősége jobban romlik prozódia-módosítás esetén.) A szintetizált mondatok mellett felvettük azok természetes változatát is, a beszédkorpusz hangját adó ugyanazon bemondó közreműködésével.

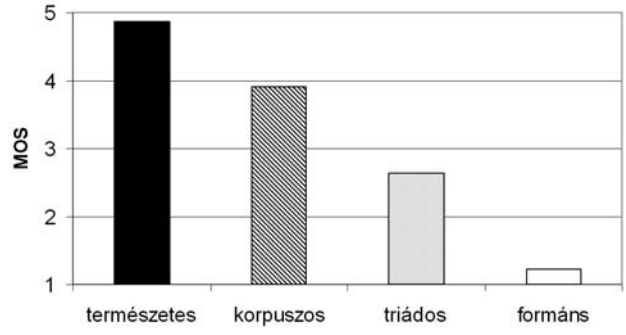
Korábbi tapasztalataink alapján egy 30 perces meghallgatásos teszt folyamán az átlagos motivációjú tesztelő elveszíti érdeklődését. Ennek elkerülésére egy körülbelül 10 perces tesztet állítottunk össze, melyben kizárólag a szintetizált beszéd minőségét kellett értékelniük a tesztelőknek. A tesztben szereplő szövegek tartalmát a kísérleti korpusz-alapú szintetizátor által megcélzott időjárás-jelentés felolvasáshoz igazítottuk. A tesztanyag tíz időjárás-jelentésekből származó mondatot tartalmazott *(5. táblázat)*. A mondatokat egy időjárásjelentés-portálról véletlenszerűen választottuk. Ugyanakkor a portálon megjelent korábbi anyagokat felhasználtunk a kísérleti szintetizátor beszédkorpuszának összeállításához, így azok stílusa „ismerős” volt a szintetizátor számára. Egy tesztelőnek összesen 40 mondatot kellett meghallgatnia (ebből 10 mondat természetes, 10 a Multivox szintetizátorral, 10 a Profivox szintetizátorral és 10 a kísérleti korpusz-alapú szintetizátorral lett előállítva).

A tesztet az Interneten keresztül, egy web-es felület segítségével kellett elvégezni. Ez lehetővé tette nagyszámú tesztelő részvételét. A 248 tesztelő nagy része egyetemi hallgatók közül került ki. A résztvevők átlagkora 22,9 év (a legfiatalabb 12, a legidősebb 63 éves volt). A kiértékelésre került adatok 185 férfitől és 36 nőtől származnak. A tesztelők 15%-a (27 fő) által adott értékelést – a későbbiekben részletezett okok miatt – nem használtuk fel. A 40 mondat meghallgatása előtt egy, a tesztben nem szereplő (de szintén időjárás-jelentésből származó) mondatot kellett a három szintetizátorral előállítva, illetve természetes változatban meghallgatni. Így minden egyes tesztelő kialakíthatta saját szubjektív rangsorát a későbbi teszteléshez. Egy ilyen előzetes „ismerkedési” fázis általában része a MOS-típusú teszteknek. A teszt folyamán minden mondatot csak egyszer lehetett meghallgatni. Ez csökkentette a teszt elvégzéséhez szükséges időt. A tesztelők által a meghallgatáshoz használt eszközök minőségét érthető okok miatt nem tudtuk szabályozni. Ugyanakkor a teszt elején kértünk erre vonatkozó adatokat. Ezek szerint a tesztelők többsége átlagos, otthoni minőségű eszközön végezte a tesztet, csendes környezetben.

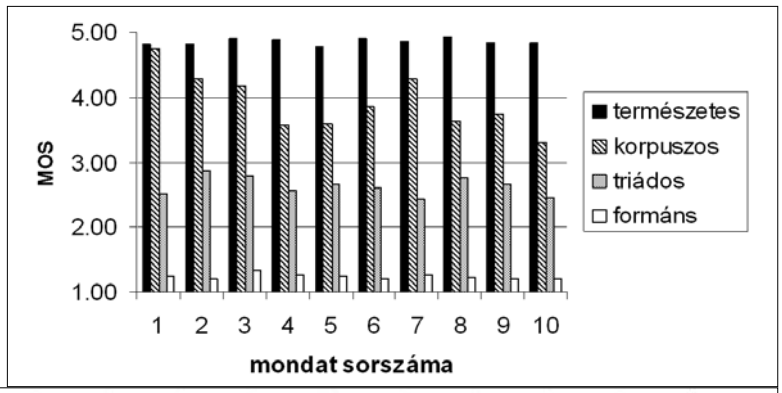
A kiértékelésből kizártuk azokat a tesztelőket, akik a tízből legalább két esetben közepes (3), vagy rosszabb minősítést adtak a természetes beszédből származó mondatokra. Feltételeztük, hogy ezek a tesztelők vagy komolytalanul „össze-vissza” válaszoltak, vagy technikai problémák miatt akadozással hallhatták az egyes felvételeket. Az utóbbi problémával mi is találkoztunk előzetes tesztjeink során. A probléma lassabb Internetkapcsolat esetén jelentkezett, és a hangfájl letöltés akadozása miatt keletkezett. Ezért a 22 kHz-es monó hangfájlokat egy 32 kbps sebességű MPEG1-LIII kódolóval tömörítettük, változó bitsebességű módban. Az esetek többségében sikerült megszűntetni a lejátszás korábbi akadozását. Stúdió minőségű fejhallgatóval végzett informális meghallgatás során nem találtunk észrevehető különbséget az eredeti és a tömörített felvételek hangzása között. A fentieken kívül azokat a tesztelőket is kizártuk, akiknek az eredményét webszerver túlterhelés miatt nem tudtuk helyesen rögzíteni. Összességében a tesztelők 15%-át zártuk ki.

A *4. ábra* és a *3. táblázat* mutatja a teszt összesített eredményét. Az eredményekből jól érzékelhető, hogy a beszédszintetizátorok generációváltása minden esetben jelentős minőségjavulást eredményezett, ugyanakkor a jelenlegi legjobb rendszerek sem érik el a teljesen természetes hangzást. A generációk közötti ugrásszerű javulás a rendszerek szélesebb-körű használatát, új alkalmazások bevezetését tette, teszi lehetővé.

4. ábra és 3. táblázat
Szubjektív minősítés átlagai
az egyes szintézis technológiákra



	természetes	korpuszos	triádos	formáns
MOS	**4,86**	**3,92**	**2,63**	**1,24**
konfidencia (α = 0,05)	**0,02**	**0,03**	**0,03**	**0,02**



5. ábra és 4. táblázat Szubjektív minősítés átlagai mondatonként. A MOS értékek konfidenciája (α=0,05) 0,04 és 0,09 közötti.

sorszám	1	2	3	4	5	6	7	8	9	10	szórás
természetes	**4,83**	**4,83**	**4,91**	**4,88**	**4,79**	**4,90**	**4,86**	**4,92**	**4,84**	**4,84**	0,04
korpuszos	**4,75**	**4,28**	**4,16**	**3,56**	**3,59**	**3,85**	**4,29**	**3,63**	**3,75**	**3,30**	0,44
triádos	**2,52**	**2,88**	**2,79**	**2,56**	**2,66**	**2,60**	**2,44**	**2,76**	**2,65**	**2,46**	0,15
formáns	**1,25**	**1,20**	**1,33**	**1,26**	**1,25**	**1,20**	**1,26**	**1,21**	**1,21**	**1,20**	0,04

Az *5. ábra* és a *4. táblázat* mutatja a mondatonkénti átlagos minősítéseket. A táblázatban szereplő szórás értékeket összehasonlítva látható, hogy a korpusz-alapú, elemkiválasztásos szintetizátor minőségének a legnagyobb az ingadozása.

A rendszer legjobban sikerült mondata elérte a természetes minőséget, a legrosszabb mondat pedig 1,5 jeggyel rosszabb minősítést kapott. Ugyanakkor az összes mondat jelentősen meghaladja a korábbi rendszerek minőségét. Ez az ingadozás egyrészt a technológia velejárója, másrészt a kísérleti rendszer hiányosságaiból, például a leegyszerűsített prozódia-előállításból származik. Várakozásaink szerint a rendszer továbbfejlesztett változatában a rosszabban hangzó mondatok minősége javulni fog, csökkentve a minőség ingadozását.

A következőkben azt vizsgáltuk, hogyan függ össze a korpuszos szintetizátorral előállított mondatok minősége az azokban szereplő vágási pontok számával. Az *5. táblázatban* megadjuk a tesztben szereplő mondatokat, szóközzel jelölve a vágási pontokat, aláhúzással jelölve a beszédszüneteket.

A *6. táblázat* számszerűleg is összesíti a vágási pontokat. A mondatokat a bennük szereplő vágási pontok száma szerint rendeztük növekvő sorrendbe. A legjobb minősítést a legkevesebb, azaz három vágási pontot tartalmazó mondat érte el, míg a legrosszabb minősége a legtöbb, 24 vágási pontot tartalmazó mondatnak volt. Ugyanakkor a minőség nem függ konzisztensen a vágási pontok számától. Például a 7. (12 vágási pontot tartalmazó) mondat minősége a második legjobb, megelőzve sok kevesebb vágást tartalmazó

1	éjszakatöbbfelépárássáválika levegő_foltokban ködis képződhet
2	a párás_ néhol ködösreggeltkövetőentöbbórára kisütanap
3	akövetkező napokbanfolytatódik az igazitéli időjárás
4	felhős égboltra _ jelentős esőzések kialakulására kellszámítani
5	azészakkeleti szél helyenként megélénkül _ csütörtökön a tiszántúlon néhol megerősödik
6	aszél mérsékelt északkeleti lesz _a hőmérséklet tizenegy _ tizenhat fok körülalakul
7	budapesten eleinte többnyireerősen felhősleszazég _ majd csökkena felhőzet_ és csütörtökön néhány órára kisütanap
8	péntektől ismét csapadékosra fordulazidő _ többfelé várható havaseső _ havazás _és egyúttal lehűlés is kezdődik
9	csütörtökön északkeleten többórárakisüta nap_ délnyugaton azonban méggyakran leszerősen felhősazég_ ésott helyenként kisebbeső _zápor továbbra is valószínű
10	akövetkezőnapokban a cs ila gá sz a ti tél az átlagosnál jóvalenyhébb idővelin d ít _ sok felhőreszámíthatunk_ és főleg szerdán kell számítani sokfelé esőre

5. táblázat A teszteléshez használt mondatok tartalma. A szóközök a korpuszos szintetizátor vágási pontjait jelölik, míg az aláhúzások a beszédszüneteket.

mondat sorszáma	1	2	3	4	5	6	7	8	9	10
MOS (korpuszos)	**4,75**	**4,28**	**4,16**	**3,56**	**3,59**	**3,85**	**4,29**	**3,63**	**3,75**	**3,30**
vágási pontok száma	**3**	**4**	**4**	**6**	**9**	**10**	**12**	**14**	**15**	**24**
szavak száma	10	11	8	7	10	12	9	15	25	22
elemekből összefűzött szavak száma	0	0	0	0	0	0	0	0	0	2

6. táblázat Szubjektív minősítés (-0,68-as) korrelációja a vágási pontok számával, korpusz alapú elem-összefűzéses kísérleti beszédszintetizátorban.

mondatot. A vágási pontok száma és a minősítés között -0,68 a korreláció, ami szintén a jelentős, de nem teljesen konzisztens összefüggésre utal. A 6. táblázatból az is látszik, hogy csak egy mondatban került sor hangszintű elemek használatára. A mondatokat meghallgatva megállapítható, hogy minőségi problémák leginkább a helytelen, nem illeszkedő prozódiából adódnak.

Ugyanakkor látható, hogy az egyszerű modell is tízből négy-öt esetben jónak minősített prozódiát ad. A prozódia-simítás, illetve a pozíció alapú megközelítés helyett komplexebb prozódiai modell megvalósításával a probléma várhatóan csökkenthető.

7. Összefoglalás

A beszédszintetizátorok generációváltása minden esetben jelentős minőségjavulást eredményezett, ugyanakkor a jelenlegi rendszerek sem érik el a teljesen természetes hangzást.

A generációk közötti ugrásszerű javulás a rendszerek szélesebb körű használatát, új alkalmazások bevezetését tette, teszi lehetővé. Látható, hogy a generációk közötti váltás nem a teljes rendszert, hanem csak egyes részeinek lecserélését érinti. Míg a korábbi rendszerek viszonylag függetlenül fejlődtek a beszédfelismeréstől, a legutóbbi, korpusz alapú, elemkiválasztásos rendszer nagymértékben támaszkodik az automatikus beszédfelismerés technológiájára.

A cikkben bemutatott kísérleti korpusz alapú elemkiválasztásos beszédszintetizátor már jelen állapotában is meghaladja a korábbi rendszerek minőségét, korlátozott tematikájú területen. A rendszer minősége ugyanakkor egyenetlen, de ez várhatóan további fejlesztéssel csökkenthető.

A korlátozott tematikájú rendszer fejlesztése során nyert tapasztalatok alapján egy általános elemkiválasztásos szintetizátor készítése is elérhető távlatba került.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani a szubjektív kiértékelésben résztvevő nagyszámú tesztelőnek. Külön köszönet illeti Bartalis Mátyást a webes tesztfelület elkészítéséért és Mihajlik Pétert a magyar nyelvű beszédfelismerő eszközök használatához nyújtott segítségéért.

A kutatást az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Irodalom

[1] Olaszy, G., G. Gordos, G. Németh:
„The MULTIVOX multilingual text-to-speech converter”, In: G. Bailly, C. Benoit and T. Sawallis (eds.): Talking machines: Theories, Models and Applications, Elsevier, 1992. pp.385–411.

[2] Olaszy Gábor:
„A Korpusz alapú beszédszintézis nyelvi, fonetikai kérdései”, jelen számban.

[3] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G.:
„PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications”, International Journal of Speech Technology, Vol. 3, Numbers 3/4, December 2000, pp.201–216.

[4] Nagy András, Pesti Péter, Németh Géza, Bőhm Tamás:
„Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései”, Híradástechnika, 2005/1, pp.18–24.

[5] Olaszy Gábor:
Beszédadatbázisok készítése gépi beszédelőállításhoz. Beszédkutatás`99 (Szerk.: Gósy Mária – MTA Nyelvtudományi Intézet), Budapest 1999. pp.68–89.

[6] Olaszi Péter:
„Magyar nyelvű szöveg-beszéd átalakítás: nyelvi modellek, algoritmusok és megvalósításuk”, (Ph.D. értekezés), BME 2002.

[7] Paul Boersma:
„Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”, IFA Proceedings 17: pp.97–110.

[8] Mihajlik P., Révész T., Tatai P.:
„Phonetic Transcription in Automatic Speech Recognition”, Acta Linguistica Hungarica, Vol. 49 (3-4), 2002. pp.407–425.

[9] ITU-T:
„A method for subjective performance assessment of the quality of speech voice output devices”, Draft ITU-T, Recom. P.85, COM 12-R 6, 1993.

[10] Alvarez, Y. and M. Huckvale:
„The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems”, Proc. International Conference on Spoken Language Processing (ICSLP), Denver 2002, Vol. 1, pp.329–332.

Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére

TAKÁCS GYÖRGY, TIHANYI ATTILA, BÁRDI TAMÁS, FELDHOFFER GERGELY, SRANCSIK BÁLINT

Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
{takacs.gyorgy, tihanyia, bardi, flugi, sraba}@itk.ppke.hu

Lektorált

Kulcsszavak: audiovizuális beszédfeldolgozás, fej animáció, multimodális kommunikációs, szájrólolvasás

Siketek kommunikációs segédeszközeként egy beszédjelet közvetlenül szájmozgás-képpé átalakító rendszert fejlesztettünk. Az előzetes vizsgálati eredmények alapján, képzett jeltolmácsokkal készítettünk kép- és hangfelvételeket. Az MPEG-4 szabványnak megfelelő egységet használtunk fejmodellnek a beszédszervek mozgásának megjelenítésére. Egy neurális hálózat számolja ki a jellegzetes pontok főkomponens súlytényező értékeit a beszédjelből. A fejmodell vezérlő paramétereit a rendszer a főkomponens súlyértékekből származtatja. A rendszer terveink szerint egy alkalmas mobiltelefonon is futtatható. A tesztvizsgálat során siket személyek a szavak közel 50%-át értették meg helyesen a fejmodell által megjelenített mozgókép alapján.

1. Bevezetés

Siket emberekben hosszú gyakorlás után fantasztikus szintre fejlődik ki a beszéd megértése pusztán a szájmozgást nézve. Az volt a tervünk, hogy erre alapozott kommunikációs segédeszközt készítünk siket felhasználók számára, amely pusztán a szájrólolvasáson alapul és egy alkalmas mobiltelefon készüléken megvalósítható. Az általunk kifejlesztett rendszerben egy beszélő fej fontos részeit jelenítjük meg a színes grafikus kijelzőn. A mozgó fej vezérlő paramétereit közvetlenül a beszédjelből származtatott jellemzők alapján számoljuk ki. Tisztában vagyunk azzal, hogy az emberi beszédfolyamatnak ez csak egy részleges megjelenítése és azzal is, hogy elvéből fakadóan is hordoz hibákat. Arra számítunk, hogy korlátai ellenére a siketek hasznos kommunikációs segédeszközhöz juthatnak rendszerünkkel és természetes módon akár telefonon keresztül is szót érthetnek a hallók többségével. Reményeink szerint ezzel is lebontható egy akadály, ráadásul mindössze olyan hétköznapi eszközzel, mint egy megfelelő kategóriájú mobiltelefon. Természetesen rendszerünk nagyban épít a siketek kifinomult képességeire és a közvetlen kommunikációban kialakult folyamatos kiegészítő és hibajavító mechanizmusaira

Jelfeldolgozási szempontból a rendszer sarkalatos eleme, hogy időkeretenként meghatározott folyamatos jellegű beszédjellemzőkből folyamatos képjellemzőket számol. Az eddig ismert megoldások leképezték a folyamatos beszédfolyamatot diszkrét nyelvi elemek (fonémák, vizémák) halmazára. Egy második lépésben pedig a diszkrét elemek halmazát alakították át mozgó fejjé. Nagy előnye a mi közvetlen rendszerünknek, hogy eredendően megőrzi a beszédfolyamat eredeti időbeli és energia-szerkezetét. Ezáltal a természetes beszédritmus eleve megőrződik. További előnye, hogy egy mobiltelefon korlátozott processzorteljesítménye, memóriakapacitása mellett is megvalósítható, és még ígéretesebb jellemzője, hogy elvileg nyelvfüggetlen.

Új ötlet a rendszerben, hogy a folyamatot nem átlagos beszélők jeleivel tanítottuk, hanem olyan hang és kép adatbázissal, amelyet képzett jeltolmácsok felvételeiből állítottunk össze. Az ő artikulációs stílusuk és dinamikájuk alkalmazkodott a siketek szájrólolvasási igényeihez.

Az irodalomban ismert szájrólolvasáshoz kapcsolódó mozgó fej alkalmazások érdekes csoportja foglalkozik azzal, hogy többletinformációt adjon a hallott beszédhez például zajos környezetben vagy nagyothallók számára [1,2,3]. A hallott és egyben látott beszédfolyamatban a szuperadditív megértés nagyobb, mint a külön modalitásban megértett elemek összege. Fontos kérdés, hogy hol és hogyan összegződnek az egyes modalitásból származó információ elemek. A mi alkalmazásunkban azonban csak a látás alapú beszédérzékelésre összpontosítottunk, mivel a célközösségben a hallás gyakorlatilag teljesen hiányzik.

A szájmozgás dinamikája és természetessége tűnik az alkalmazás kritikus elemének. Számos közlemény számol be arról, hogy milyen bonyolult eljárásokkal érik el a beszélő fej modell megfelelő dinamikáját és természetességét [4,5,6].

Mi ezt pusztán azzal kívántuk elérni, hogy különös figyelemmel választottuk ki az adatbázisba bevont beszélő személyeket. Ezek tehát nem az átlagos népességet, hanem a siketek számára legjobban érthető beszélőket reprezentálják. Mi ezzel a trükkel oldottuk meg a nagyobb szájmozgás dinamikát igénylő követelményeket.

2. Adatbázistervezés és összehasonlítás

2.1. Előzetes szájrólolvasási mérések

A kezdeti vizsgálatokban mértük a siketek szájrólolvasási képességeit, feltérképeztük mindennapi kommunikációs problémáik lényegét. A részleteket korábbi cikkünkben ismertettük [13], itt most csak a végkövetkeztetéseket foglaljuk össze.

Legfontosabb végkövetkeztetéseink egyike volt, hogy a szájról olvasott beszéd érthetősége nagyon függ az artikuláció minőségétől. A szájrólolvasás sokkal nagyobb figyelmet igényel, mint a beszéd megértése hallás útján. A teljes beszédfolyamatról csak részleges információt ad, ezért a tévesztések eleve gyakoribbak. Az olyan artikuláció, amely eleve kiemeli a megkülönböztető jegyeket, valamint a lassú beszédtempó nagyon sokat segít a helyes megértésben. A hallók között messze legjobban teljesítik ezeket a követelményeket a képzett jeltolmácsok. Ãk napi kapcsolatban állnak a siketekkel és ezért eleve alkalmazkodik artikulációjuk a szájrólolvasás igényeihez. Ezért határoztuk el, hogy tanító adatbázisunkat jeltolmácsok kép- és hangfelvételeiből állítjuk össze, még akkor is, ha a végső használatkor bárkinek a hangja szolgálhat jelbemenetként.

Megtanultuk az előzetes kísérletek során azt is, hogy a siketeknek komoly nehézségeik vannak a természetes nyelv komplikált nyelvtani szabályaival. Elektronikus leveleiket, SMS üzeneteiket is elemezve látszik, hogy ugyanez megnyilvánul írott kommunikációjukban is. Amikor az érthetőséget teljes mondatok, rövid közlendők formájában adott nyelvi egységekkel próbáltuk mérni, akkor tapasztaltuk, hogy nem képesek a teljes üzenet pontos, szó szerinti visszaadására, hanem csak a legfontosabb üzenetelemek maradnak meg emlékezetükben. Sokszor csak az a kulcselem, amelyre az előzetes információk alapján a figyelmük középpontjába kerül. A ragokkal, toldalékokkal sem nagyon foglalkoznak. Konkrét nevek, személyes névmások fontosabbak számukra.

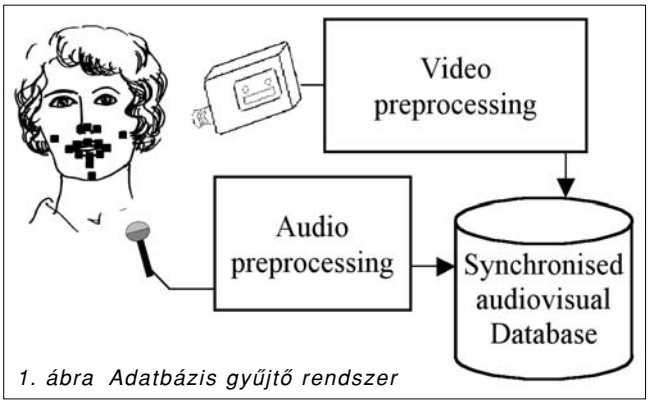
Egy hirtelen témaváltás is igen nehezen követhető számukra. Ennélfogva az érthetőségvizsgálatok szokásos szövegei és módszerei eleve nem használhatók esetükben. A tényleges érthetőség mérése érdekében ezek szövegösszefüggéstől lehetőleg mentes szöveget használnak, ami teljesen idegen a siketek kommunikációs stratégiájától. Emiatt speciális szövegű adatbázist alakítottunk ki mind a tanító, mind a tesztelő anyaghoz.

Az előzetes vizsgálatok fontos kérdése volt, hogy két vagy három dimenziós fejmodellt kell-e megvalósítani, és hogy mennyire fontos a harmadik (mélység) dimenzió a szájrólolvasás során. Ennek eldöntésére természetes beszélő személyek videofelvételeit mutattuk siketeknek olyan torzítások után, amelyek a mélységinformációt csökkentették. Az egyik esetben csak a kék színösszetevőt tartottuk meg és a piros és zöld színösszetevőket kivettük a képből. További felvételeknél pedig csak fehér vagy fekete képpontok maradtak az eredeti képből egy alkalmas küszöbszintet választva. Meglepő módon ezek a torzítások alig csökkentették a szájrólolvasás pontosságát, pedig a mélységinformációt kiölték a felvételekből.

További kísérleteinkben arra kerestünk választ, hogy a jobb mobiltelefonoknál szokásos képernyő méret és felbontás elegendő-e a szájrólolvasáshoz. Amennyiben a kijelzőn megjelenő kép a szájat és környékét mutatja (a beszédinformációt hordozó legfontosabb részeket), akkor ez a méret és felbontás eredeti videofelvételek esetén elegendő a gyakorlatilag teljes megértéshez.

2.2. Felvételek

Az adatbázis nem más, mint különböző bemondók összerendezett hangfelvételeinek és képfelvételeinek rendszere. A felvételek jelét azonos időkeretekben összeszinkronizálva dolgoztuk fel *(1. ábra)*. A bemondók fejét puha korlátokkal rögzítettük, hogy a fej ingatását megakadályozzuk. Az egyes pontokat abszolút koordinátáikkal jellemezhetők.



1. ábra Adatbázis gyűjtő rendszer

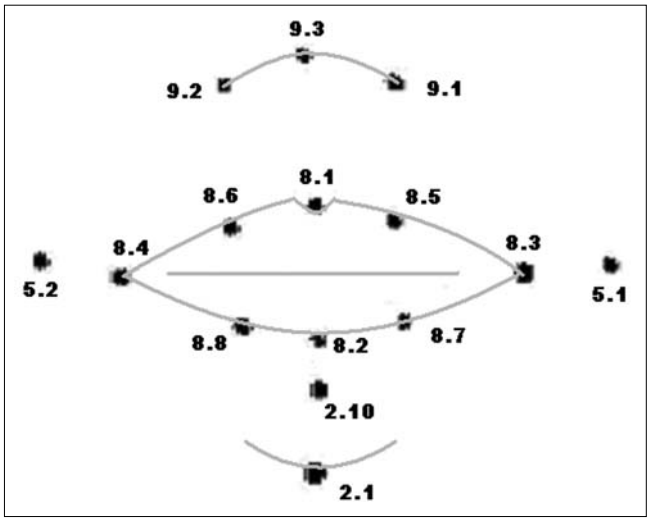
Jelen állapotában a rendszer bemondófüggő, de már dolgozunk a személytől független megoldáson is.

Az MPEG-4 szabvány az emberi arcot 86 jellemző ponttal (Feature Point, FP) írja le. Előzetes kísérleteink alapján ezekből 15-öt választottunk ki a szájnak és környezetének leírására. A felvételek során ezeket a pontokat könnyen lemosható és egészségre nem ártalmas festékkel jelöltük meg a bemondók arcán. A beszédfolyamat képének leírása az MPEG-4 szabvány szerinti jellemző pontokkal több szempontból is előnyös. Egyrészt a száj és arc mozgásának tömör és elég pontos leírására alkalmasak az FP koordináták, másrészt a bevált szabványos fejmodellek alkalmazhatók ezekkel a pontokkal vezérelve, így az igen összetett modellek alapvető fejlesztésére nem kellett erőinket pazarolni. Amint az előző pontban kifejtettük csak képzett jeltolmácsokkal készítettünk felvételeket.

A felvételekhez egyszerű kamerákat használtunk: 720x576 pontos felbontással, másodpercenként 25 képpel, PAL szabvány szerint. Ez azt jelenti, hogy 40 ms hosszú időablakokban készülhettek az összeszinkronizált kép- és hangelemzések. A felvételek a szájat és környékét rögzítették annak érdekében, hogy a kiválasztott jellemző pontok helyzete minél kisebb hibával meghatározható legyen. A fej többi részét (bár a szem környéke, vagy akár a hozzáfűzött tekintet is hordoz tartalmi információt) nem vontuk bele vizsgálatainkba. A képfelvételeket ezután emberi beavatkozás nélkül dolgoztuk fel. A képjelet a kontraszt, a fényesség és telítettség tekintetében úgy torzítottuk, hogy a jellemző sárga pontok minél jobban kiemelődjenek. A sárga pontokat végül az RGB komponensek kompará-

lásával detektáltuk. A binarizált képen először dilatációs műveleteket végeztünk, hogy biztosan összefüggő képpont halmazt nyerjünk, majd lépésenként kívülről eróziós folyamattal szedtünk le képpontokat, míg egyetlen pixel maradt, amit a jellemző pont közepének tekintettünk. Ez az automatikus eljárás legfeljebb 1-2 pixel eltérést eredményez a manuálisan kiválasztott középponthoz képest.

Tekintettel arra, hogy az egyes FP jellemző pontok vízszintesen 40-60, függőlegesen 80-140 pixel tartományban mozognak, az FP meghatározás fenti hibája elfogadható. A koordinátarendszert úgy választottuk meg, hogy középpontja az orr két oldalára helyezett (9.1 és 9.2) pontok között középen legyen, mivel ezek a pontok mozognak a 15 közül legkevésbé *(2. ábra)*.

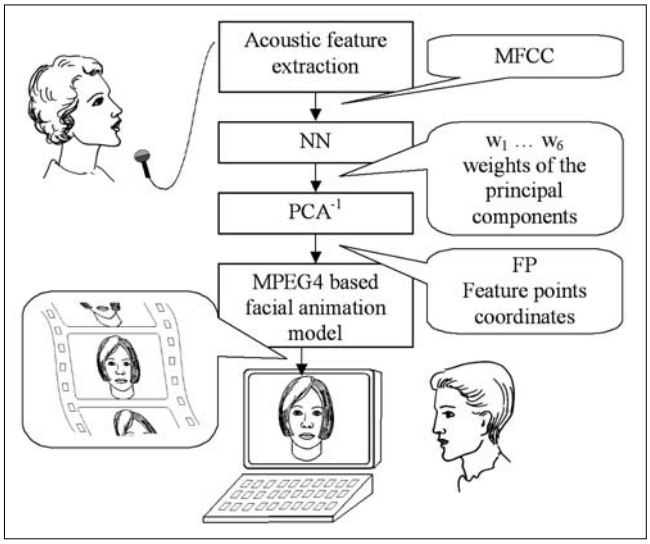


2. ábra Az MPEG-4 jellemző pontok kiválasztott részhalmaza a száj körül

A beszédjelet egy hangcsatornában rögzítettük 48 kHz mintavételezéssel, 16 bites mintákkal.

A tanító és tesztelő adatbázis szövegét a 2.1 pontban leírt követelmények szerint választottuk ki. Eszerint a felvételek kétjegyű számokat, hónapok neveit, a hét napjait tartalmazták.

3. ábra A beszéd-szájmozgás átalakító rendszer elemei



3. A beszédjel átalakítása szájmozgás-képpé

A fejlesztés állapotában a rendszer lényegében egy személyi számítógépen futó programrendszer. A *3. ábra* szerint itt az alapelemek feladatait és kapcsolódását tekintjük át. Az egyes elemek részleteit a 3.1-3.4 pontok fejtik ki

A mintavételezett beszédjelen minden 40 ms keretben meghatároztuk a mel skála szerinti kepsztrum együttható vektort (Mel-Frequency Cepstrum Coefficients, MFCC). Ezeket a jellemző vektorokat vezettük a neurális hálózat (NN) bemenetére, amely a kimenetein kiadja a szájmozgás pillanatnyi állapotát tömörítetten leíró súlytényező vektort $[w_1,...w_6]$. A főkomponens elemzés (PCA) inverz műveletével nyerjük a fejmodell vezérléséhez ténylegesen szükséges FP koordináta értékeket. Ez egy lineáris kombinációs műveletet jelent csupán. Az FP koordinátákat meghatározzuk minden időkeretre. Erre láthatunk egy példát az *5. ábrán*.

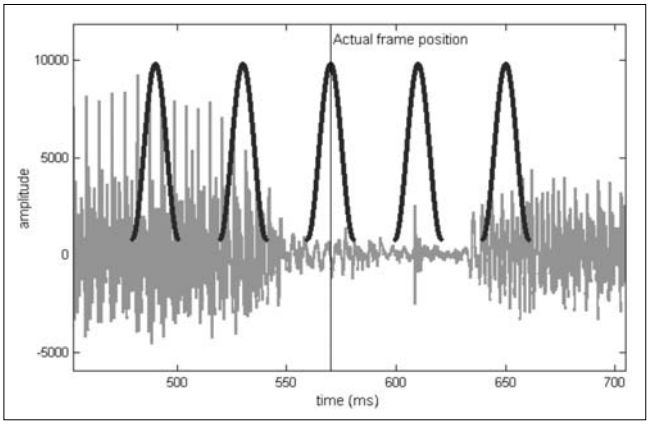
Rendszerünk utolsó eleme a nyílt forráskódú LUCIA beszélő fej rendszernek egy enyhén módosított változata. A modellt az FP koordinátákkal vezéreljük és a mozgó kép megjelenik a kijelzőn. A részletek a 3.4 fejezetben találhatók.

3.1. Akusztikai lényegkiemelés

A bejövő beszédjelen először egy magasemelő szűrési műveletet hajtunk végre $H(z) = 1\text{-}0.983z^{-1}$ karakterisztikával. Ezután 21.33 ms hosszúságú Hamming-ablakkal súlyozzuk a jelet. Az ablakban lévő jelből 16 elemű mel-frekvenciás kepsztrum együttható vektort számolunk.

A koartikuláció jelenségének a beszédfolyamat képi ábrázolásánál legalább akkora jelentősége van, mint a hangjelek feldolgozásakor. A beszédszervek képe szempontjából vannak domináns és változó fonémák. A domináns fonémák kifejezetten megszabják a száj és környezete képét viszont a változó típusok képét a környező domináns fonémák nagyban befolyásolják. Ebből fakadóan a beszédjelből a beszédszervek képét becslő algoritmusnak a szomszédos kereteket is felölelő környezetre is tekintettel kell lennie.

4. ábra Egy fonémaátmenet jellemzése öt egymás utáni keret alapján



A siket partnerek számára a lassabb beszédtempó vezet eredményre. Gyakorlott jeltolmácsok a beszédhangok tisztafázisú részét világosan és kiemelve képzik. Másodpercenként 5-10 beszédhangot ejtve és 40 ms hosszú elemzési időkereteket tekintve 5 elemzési ablak egyike bizonyosan ráesik a beszédfolyamat legalább egy domináns fonémájára *(4. ábra)*.

A neurális hálózat bemenetére tehát mindig 5 egymás utáni elemzési ablak kepsztrum vektora kerül.

3.2. A neurális hálózat

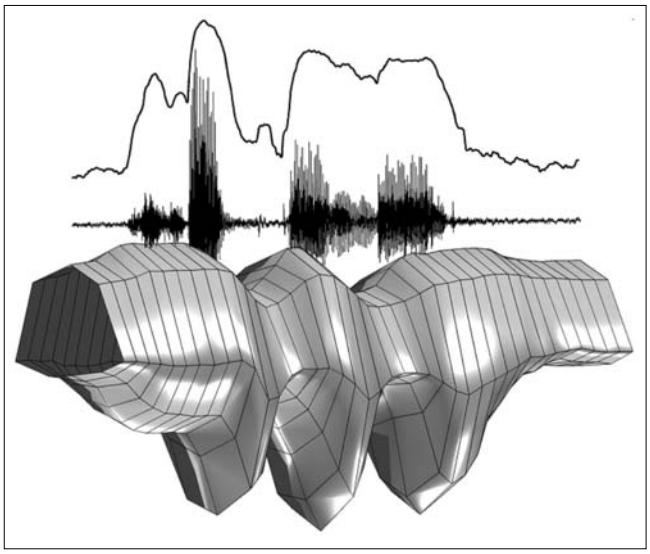
A visszacsatolt neurális hálózatot a hagyományos hibajel visszaterjedéses módszerrel tanítottuk, azzal a programmal, amelyet David Anguita fejlesztett ki és tett közzé [8].

A hálózat három rétegben 80 csomópontot tartalmaz. A bemeneti réteg fogadja 80 ponton 5 egymás utáni időkeret 16-16 MFCC értékét. A rejtett réteg 40 csomópontot tartalmaz. A kimenő réteg 6 csomóponton szolgáltatja a 6 főkomponens súlyértékét, amelyekből előállítható a 15 jellemző pont (FP) x-y koordináta értéke a középső időkeretben.

A tanító adatbázis 5450 időkeretet tartalmazott. A hálózat tanítását 100.000 ciklusban végeztük. A neurális hálózat modell a bemeneti és kimeneti változók értékeit a -1, 1 értéktartományba normálta. Az MFCC és PCA változókat mind ebbe a tartományba transzformáltuk lineárisan az MFCC vektor energia összetevőjének kivételével.

A már betanított neurális hálózat programja igen gyorsan futtatható, mivel az egész adatbázist képviseli a hálózat súlytényező vektor, amely mindössze 3440 elemből áll. A hálózat kimeneti értékeinek valós idejű számolásához tehát egy alkalmas mobiltelefon erőforrásai elegendőek.

5. ábra
A 8.1-8.8 jelű jellemző pontok x-y koordinátái az idő függvényében a „szeptember” szó kiejtésekor. A felső folyamatos vonal a keretenkénti energiát ábrázolja dB skálán, a középső görbe a hullámformát mutatja. Az alsó ábrán látható felület az ajakkontúrokat mutatja.



3.3. Főkomponens analízis (PCA)

A képfelvétel minden időkeretében 15 jellemző pont írja le a száj és környékének pillanatnyi alakját. A kétdimenziós ábrázolás alapján ez egy 30 dimenziós térben egy ponttal jellemezhető. A rendszer tanítása sokkal hatékonyabbá vált azáltal, hogy a 30 dimenziót 6 dimenziós rendszerré tömörítettük.

A dimenzió redukció végrehajtására a főkomponens elemzés módszerét (Principal Component Analysis, PCA) alkalmaztuk. Ez felfogható mozgáskomponensek szerinti felbontásra, amint ezt a *6. ábra* mutatja. Az első 6 PCA vektort választottuk a száj és környékének leírására az alábbi egyenlet szerint

$$w_{1\ldots6} = P^{-1}B\Big|_{p_1^{-1}\times..\times p_6^{-1}} \tag{1}$$

ahol *P* jelöli a PCA vektorok (30x30) méretű sajátérték vektorát, *B* a 30 dimenziós vektor készlet, *c* pedig a választott origó, amely a zárt ajkakkal semleges arc súlytényezőinek 0 értékét jelenti. Ez az adattömörítés mindössze 1-3% hibát eredményezett, ami az adott megjelenítő eszközön a jellemző pontok 1-2 pixeles változását eredményezi akár x, akár y koordináta szerint nézve. Ez teljesen elfogadható közelítés. Mivel a hálózat tanításához használt *w* súlytényező 0 értéke a semleges archoz tartozik, ezért a súlytényező előjele is egy nagyon fontos információt hordoz: megmutatja, hogy a pont merre mozdul el.

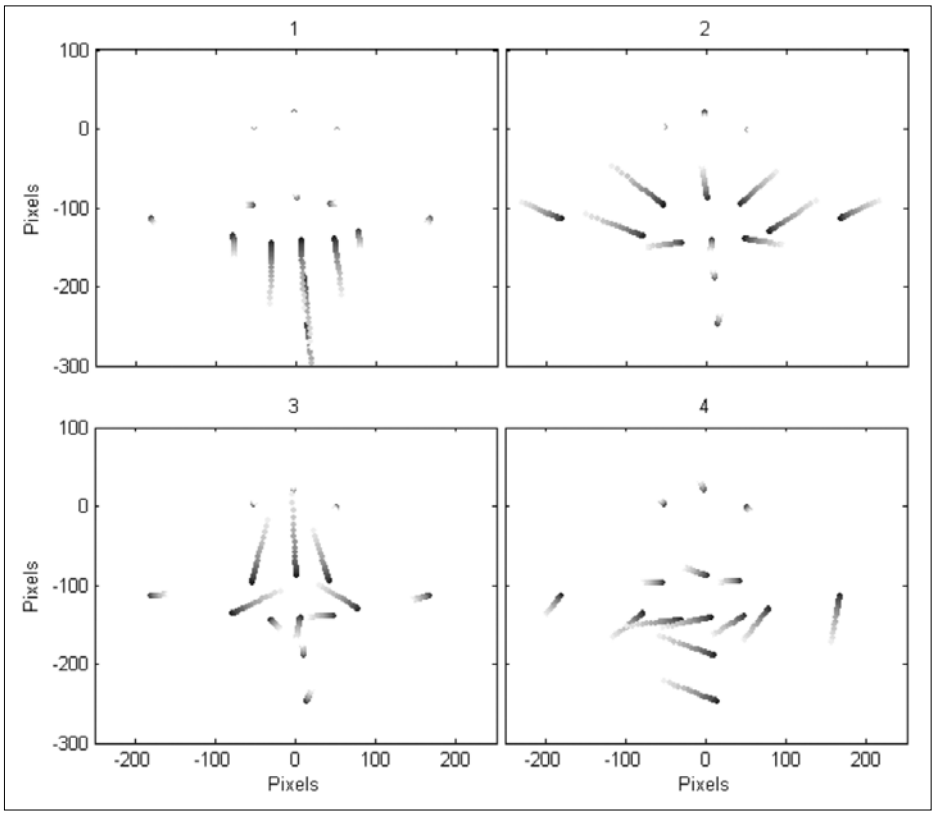
A betanított hálózat kimenő értéke egy 6-dimenziós térben jelenik meg. Ebből a jellemző pontok koordinátái a következő egyenlet segítségével határozhatók meg.

$$\overline{B}_k = \left(\underline{w}_k + \underline{c}\right)\cdot P \tag{2}$$

Mivel *P* értékét a tanítás során határozzuk meg, ezért ez a művelet mindössze 180 szorzást igényel keretenként.

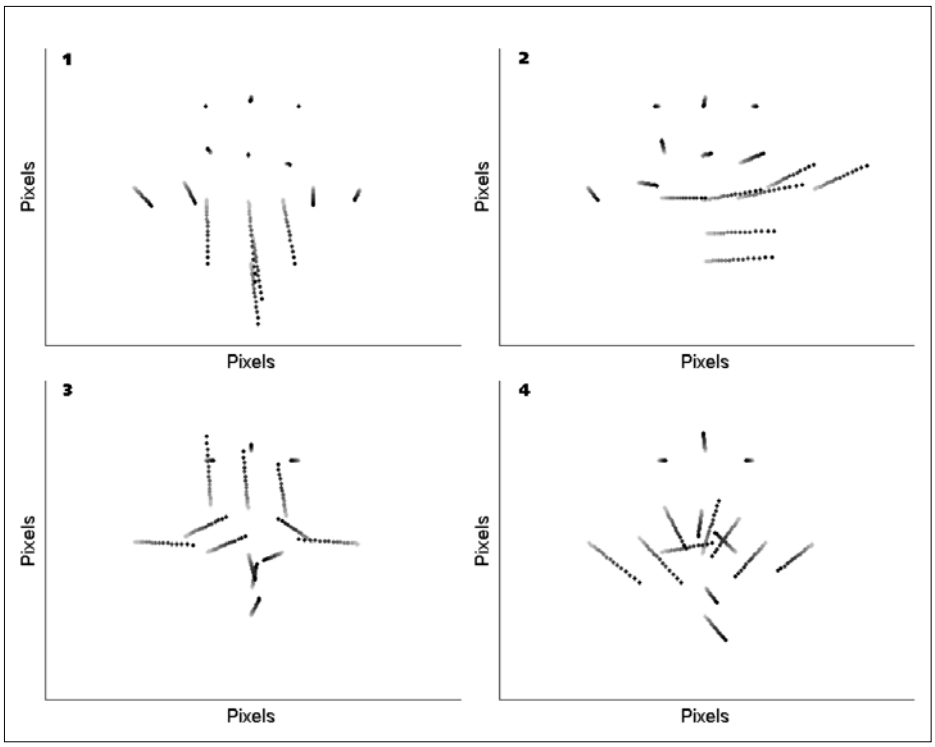
A főkomponens analízis ebben az esetben több, mint egy egyszerű mechanikus tömörítő eljárás. A PCA vektorok értékes információt hordoznak a bemondó beszédstílusáról is és a felvétel minőségéről is. A PCA vektorok – bár automatikus eljárás eredményeként adódnak – az egyes vizémák jól azonosítható megkülönböztető jegyeihez kapcsolódnak. Szépen kiolvasható ez a 6. ábrán is. Az állkapocs függőlegesen látszó mozgása adja a legerősebb PCA komponenst. A száj vízszintes széthúzása adja a második főkomponens nagy részét (erre a mozgásra kéri fel a fényképész az érintetteket azzal, hogy mondják „csíííz”). Jól megfigyelhető, hogy a harmadik főkomponens az ajakkerekítés mértékéhez kapcsolódik. Ezek miatt állítható, hogy a PCA vektorok eredendően kapcsolódnak a vizéma megkülönböztető jegyekhez.

Ezen nézőpontból a PCA vektorok dimenzió sorrendje rendelkezik kiemelt jelentőséggel. Képzett jeltolmácsoknál az első néhány főkomponens tartalmazza a vizéma megkülönböztető jegyeket. Gyakorlatlan bemondóknál azt tapasztaltuk, hogy a korrektív komponensek sorrendben megelőzik a vizémákat megkü-



6. ábra
A jellemző pontok helyzete az első, második, harmadik és negyedik főkomponens szerint kifejezve képzett jeltolmács beszédje alapján.

7. ábra
A jellemző pontok x-y koordinátái az első, második, harmadik és negyedik főkomponens értékével kifejezve gyakorlatlan bemondó felvételei alapján.



lönböztető komponenseket (korrektív komponens például az érzelmet kifejező összetevő). Ezt mutatja be a *7. ábra*, ahol a második főkomponens fejezi ki, hogy a bemondó nagyon jellemzően, ferdén mozgatja a száját. Ezért nem is érdemes felhasználni felvételét a hálózat tanítására.

3.4. Beszélő fejmodell

A szabad forráskódú programmal közzétett LUCIA fejmodell némileg módosított változatát használtuk a rendszerben. Ezt más célra, az érzelmeket is kifejező vizuális beszédmodell céljára fejlesztették Cosi és munkatársai [10].

A LUCIA modell az MPEG-4 szabványra épült. Az eredeti fejmozgató (FAP) paraméterek vizéma alapú rendszert figyelembe véve lettek kialakítva, a szájrólolvasás igényrendszerét nem vették tekintetbe a fejlesztésnél.

Ezért volt szükség némi módosításra, hogy a modell képes legyen a jellemző pont koordináták közvetlen fogadására. A közvetlen vezérlés bőrön látható pontok mozgási lehetőségeinek anatómiai alapú megkötöttségeinek finomabb figyelembe vételét követelte meg. Ennek részleteiről [8] ad tájékoztatást.

4. Kísérletek és eredmények

4.1. Előzetes vizsgálatok

Hasznosnak bizonyultak az előzetes méréseink a rendszer tökéletesítése és az adatbázis kialakítása szempontjából. Ennek során derült ki például, hogy képzett jeltolmácsokat célszerű alkalmazni a rendszer tanításánál.

Az előzetes vizsgálatok mutattak rá arra is, hogy a szavak közötti szünetekre is különös figyelmet kell fordítani. Egy küszöbszint alatti háttérzaj nem okoz gondot. Nagyobb háttérzaj óhatatlanul elkezdi mozgatni szavak között is picit a szájat és ez nagyon megzavarja a pusztán szájról olvasásra épülő beszédfelismerést.

Az előzetes vizsgálatok során a siket kísérleti személyektől összegyűlt észrevételeket, javaslatokat gondosan figyelembe vettük a rendszer tökéletesítésénél és a vizsgálati módszerek finomításánál.

4.2. Mérési módszerek és eredmények

Pusztán szájrólolvasás alapján nem lehet azonos képzési helyű és módú fonéma párokat megkülönböztetni (például baba-papa). Természetes módon az észlelő személy a szövegösszefüggésre alapozva automatikusan korrigálja vagy kiegészíti a szájról leolvasott információt. Párbeszéd esetén egy visszakérdezés tisztázni képes a többértelmű üzenetet. Vizsgálatainkban kirekesztettük a visszakérdezés lehetőségét, ezért olyan vizsgáló szöveget állítottunk össze, amely lehetőleg kizárja a kétértelműséget.

A siketek az előzetes információk alapján mindig erősen leszűkített készletű lehetséges üzenetek közül egy kiválasztására összpontosítanak a szájról olvasott beszéd megértése során. Ezt a természetes mechanizmust célszerű volt követnünk a rendszer mérése során is. Mindig megadtuk, hogy milyen zárt halmazból kell a lehetséges választ várniuk.

A vizsgálószövegben ezért kétjegyű számok, hónapok nevei, a hét napjainak nevei szerepeltek előre megadott kategória szerint.

A mérések során a modell teljes fejét, szájmozgását mutatta a kivetített mozgókép nagyméretű vetítővásznon. Természetesen hang nélkül. Így a töredékes hallással rendelkező vizsgálószemélyek sem hallhattak semmit a beszédjelből. A vizsgálati anyag véletlen rendben az alábbi eseteket tartalmazta:

A) a jeltolmács eredeti képfelvétele (hang nélkül),
B) a fejmodell mozgóképe,
ahol a 15 vezérlő paraméter (FP) koordináták értékei jeltolmács képfelvételeiből származtak (hang nélkül),
C) a fejmodell mozgóképe,
ahol a 15 vezérlő paraméter (FP) koordinátáit a rendszer a beszédjel paramétereiből számolta ki (a megjelenítés hang nélkül történt itt is).

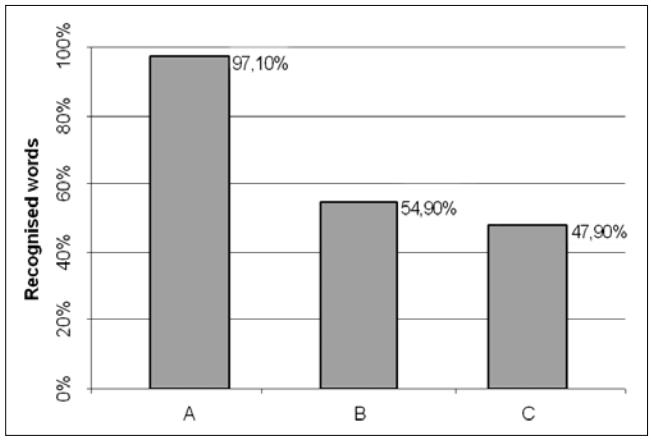
A siket vizsgálószemélyek válaszaikat írásos formában adták meg egy előkészített űrlapon. A végső eredményeket adó vizsgálat részvevői már több alkalommal rész vettek az előzetes vizsgálatokban, így mindegyikük gyakorlott mérőszemélynek volt tekinthető. A végső vizsgálat 70 szó megértését regisztrálta és mintegy 30 percig tartott. Amikor jelezték, akkor a képfelvételt megismételtük. A végső vizsgálatban 18 siket személy vett részt.

Az eredmények a *8. ábrán* láthatók.

4.3. Értékelés

A jeltolmácsok eredeti képfelvételei alapján a szavak szájrólolvasása körülbelül 3% felismerési hibát eredményezett.

A 15 FP pont koordinátáival vezérelt fejmodell, ha a vezérlő paramétereket közvetlenül a jeltolmács képfelvételein megjelölt pontok koordinátáiból származtattuk, akkor 42% felismerési hibát adott. A méréseket követő megbeszéléseken a vizsgálószemélyek olyan szóbeli megjegyzéseket tettek, hogy hiányzott bizonyos helyzetekben a modellből a nyelv képe és néha a szájtól távolabbi részek mozgása is. Emiatt a fejmodell árnyaltabb vezérlése esetleg megfontolandó.



8. ábra
A helyesen megértett szavak aránya
(A) *jeltolmács képfelvétele alapján,*
(B) *jeltolmács FP koordinátáival vezérelt fejmodell képe alapján,*
(C) *beszédjelből számolt FP koordinátákkal vezérelt fejmodell képe alapján*

A pusztán hangelemzésből számolt vezérlő paraméterekkel vezérelt fejmodell alapján mért szó érthetőség az előző esethez képest csak 7%-kal csökkent. Ez mutatja rendszerünk alapvető eredményét, azaz annak igazolt tényét, hogy a hangjelből számolt vezérlő paraméterekkel jól megközelíthető a képjelből származtatott paraméterekkel vezérelt modell felismerési aránya. Mindez épít a siket személyek kifinomult felismerési képességeire és kizárólag erre az esetre érvényes az előző megállapítás.

5. Összefoglalás

Kísérleti eredményeink igazolták, hogy lehetséges beszédjelből közvetlenül szájmozgást leíró jellemzők származtatása olyan pontossággal, ami lehetővé teszi siket személyek számára a beszéd gyakorlati hasznosságú megértését. Erre alapozva segédeszköz készíthető siketek számára, hogy megértsék csak telefonon vett beszédjelből a beszédüzenetet. A rendszer alapelemei olyan számítástechnikai erőforrással megvalósíthatók, amely rendelkezésre áll a mai legfejlettebb mobiltelefonokban.

A fejmodell további finomításától reméljük a teljes rendszer olyan fejlődését, amely révén elérhető a 20% alatti vizuális felismerési hiba, amely szint egy minden szempontból elfogadható értéket jelent. Emlékeztetünk arra, hogy a mobiltelefonok áldásaiból gyakorlatilag kirekesztett siketek közösségének ez forradalmi előre lépést jelentene jelenleg még fennálló akadályaik leküzdésében.

Köszönetnyilvánítás

A szerzők ezúton is kifejezik köszönetüket a Nemzeti Kutatási és Technológiai Hivatalnak a 472/04 szerződés keretében nyújtott támogatásáért.

A közös munka során igaz barátainkká vált siketek és jeltolmácsok lelkes közössége ösztönző példaként áll előttünk további kutatásaink során. Ezért nem csak áldozataikat köszönjük, hanem további segítségüket is kérjük.

Köszönjük Harczos Tamás kollégánk értékes ötleteit és munkáját is.

Irodalom

[1] D. W. Massaro,
Perceiving Talking Faces: From Speech Perception to a Behavioral Principle Cambridge,
Mass: MIT Press, 1998.

[2] D. W. Massaro, D. G. Stork,
"Speech Recognition and Sensory Integration,"
American Scientist, 86. 1998.

[3] M. Johansson, M. Blomberg, K. Elenius,
L. E. Hoffsten, A. Torberger,
"Phoneme recognition for the hearing impaired,"
TMH-QPSR, 2002., Vol. 44 – Fonetik, pp.109–112.

[4] K. H. Choi, J. N. Hwang,
"Constrained optim. for Audio-to Visual Conversion,"
IEEE Transactions on Signal Processing,
Vol. 52, No.6, June 2004, pp.1783–1790.

[5] R. Gutierrez-Osuna, P. K. Kakumanu, A, Esposito,
O. N. Garcia, A. Bojorquez, J. L Castillo, I. Rudomin,
„Speech-driven Facial Animation with Realistic Dynamics" IEEE Transactions on Multimedia,
Vol. 7, February 2005, pp.33–42.

[6] J. Beskow,
Talking Heads, Models and Applications for Multimodal Speech Synthesis (Doctoral Dissertation),
Stockholm, 2003.

[7] J. Ostermann,
"Animation of Synthetic Faces in MPEG-4",
Computer Animation, Philadelphia, Pennsylvania,
June 8-10, 1998., pp.49–51.

[8] Davide Anguita,
Matrix Back Propagation – An efcient implementation of the BP algorithm", Technical Report,
DIBE – University of Genova, November 1993.

[9] B. Granström, I. Karlsson, K-E Spens,
"SYNFACE – a project presentation",
Proc. of Fonetik 2002, TMH-QPSR 44: pp.93–96.

[10] Cosi P., Fusaro A., Tisato G.,
"LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", Proceedings of Eurospeech 2003,
Geneva, Switzerland, September 1, 2003.,
Vol. III, pp.2269–2272.

[12] G. Salvi:
"Truncation error and dynamics in very low latency phonetic recognition", Proc. of ISCA workshop on Non-linear Speech Processing (2003).

[13] Bárdi Tamás, Feldhoffer Gergely, Harczos Tamás,
Srancsik Bálint, Szabó Gábor Dániel:
"Audiovizuális beszéd adatbázis és alkalmazásai",
Híradástechnika, Vol. LX, 2005/10, pp.24–28.

[14] Feldhoffer Gergely, Bárdi Tamás,
Jung Gergely, Hegedűs Iván Mihály:
"Mobiltelefon alkalmazások siket felhasználóknak",
Híradástechnika, Vol. LX, 2005/10, pp.29–32.

Szöveges adatbázis tervezése rendszerüzenet generátorhoz

NÉMETH GÉZA, OLASZY GÁBOR+, BŐHM TAMÁS
BME Távközlési és Médiainformatikai Tanszék, +MTA Nyelvtudományi Intézet,
{olaszy, nemeth, bohm}@tmit.bme.hu
UGRON ZOLTÁN
EMTE Sapientia, ugron.zoltan@gmail.com

Lektorált

Kulcsszavak: beszédválaszú rendszerek, korpusz-alapú beszédszintézis, szöveges adatbázisok, beszédadatbázisok

Beszédválaszú telefonos alkalmazások elsődleges kimenetei az előre felvett beszédüzenetek (prompt-ok) – a rendszer ezeknek a bejátszásával ismerteti a felhasználóval a hívott szolgáltatással kapcsolatos választási lehetőségeit (menüpontok), visszaigazolja műveleteit stb. A promptok szövegének alacsony entrópiája miatt valószínűsíthető, hogy az emberi beszédet megközelítő minőségben előállíthatóak egy erre a célra fejlesztett beszédszintetizátor segítségével. Ennek megvalósításával kiküszöbölhetők az új üzenetek hangfelvételi nehézségei. A sikeres szintézishez szükséges, hogy a promptgenerátor adatbázisa reprezentatív legyen a várható bemeneti adatokra, azaz az előállítandó promptokat minél kevesebb beszédelemből tudja összefűzni. Cikkünkben a fejlesztés alatt álló rendszer működési elvének ismertetése után a hangadatbázis elkészítéséhez szükséges, felolvasandó szöveges állomány (szövegkorpusz) tervezési módszerét tárgyaljuk, majd bemutatjuk, hogyan vizsgáltuk meg a korpusz szövegének reprezentativitását egy független szöveggyűjtemény felhasználásával.

1. Bevezetés

Napjainkra széles körben elterjedtek az interaktív beszédválaszú rendszerek. Ezeket elsősorban automatikus ügyfélszolgálatok és információs szolgáltatások megvalósítására használják.

A kimeneti funkció (beszédüzenet) megvalósítására kétféle megoldás ismert: rögzített hangfelvételek, mint rendszerüzenetek (promptok) vagy általános szöveg-beszéd átalakító által generált beszéd (text-to-speech, TTS). Az előbbi az összes lehetséges rendszerüzenet felvételét, tárolását, és megfelelő időben való lejátszását jelenti. Ebben az esetben csak előre rögzített beszédüzeneteket használhat a rendszer (például „Önnek új elektronikus levele érkezett"). Ez nem előnyös. További hátrány, hogy ezeknek a rögzített üzeneteknek a legkisebb változtatása esetén is új hangfelvételeket kell készíteni az eredeti bemondóval.

Sokkal rugalmasabb megoldás egy általános célú TTS alkalmazása, amely tetszőleges szöveget képes érthetően felolvasni. Így elég, ha az egyes rendszerüzenetek pontos szövege futási időben alakul ki (például: „Önnek új elektronikus levele érkezett Kovács Balázstól, melynek tárgya: holnapi találkozó"). A rendszer módosítása – például új funkciók bevezetése – esetén nem kell új hangfelvételeket készíteni. Ennek a megoldásnak is van hátrányos oldala; a gyengébb hangminőség. A ma, magyar nyelven elérhető TTS-ek által előállított beszéd jól érthető, emberi beszédre emlékeztető, de gépies hangzású. Az ilyen mesterséges beszéd még nem elfogadható rendszerüzenetként a felhasználók számára. Ez a magyarázata annak, hogy a hazánkban működő interaktív beszédválaszú rendszerek szinte kivétel nélkül előre rögzített promptokat használnak.

Megfigyelhető, hogy beszédválaszú rendszerben a bemondandó promptok szókincse jóval kisebb, mint általános szövegek esetén. A mondatok szerkezete, a fogalmazás módja is kevésbé változékony. Ennek oka, hogy a szövegek egy adott témához kapcsolódnak és egy adott (ügyfél-ügyintéző párbeszédre emlékeztető) stílust követnek. Fontos megjegyezni, hogy ettől még nem korlátozott az üzenetek témája. Bármikor, akár futási időben is, megjelenhetnek korábban nem látott szavak (például új szolgáltatások, termékek bevezetésekor). Így a szókészlet nem rögzített, de jelentősen eltolódik a témába eső szavak irányába. Ugyanez igaz a mondatszerkesztésre és a szófordulatokra is.

Az interaktív beszédválaszú rendszerek adott témáját kihasználva az alkalmazáshoz illesztett olyan TTS-t lehet létrehozni, amely sokkal jobb minőségű beszédet tud előállítani, mint egy általános célú beszédszintetizátor. A rögzített promptszövegek és az általános TTS előnyeit ötvözve jó minőségű beszédet előállító, de ugyanakkor rugalmas dialógusrendszerek építhetők.

A BME TMIT Beszédkutatási Laboratóriumában egy ilyen célra felhasználható promptgenerátor fejlesztése folyik, amely egy infokommunikációs szolgáltató beszédválaszú rendszereiben használható fel. egy a fenti célnak megfelelő promptgenerátor fejlesztése folyik, amely egy infokommunikációs szolgáltató beszédválaszú rendszereiben használható fel. A munka még nem zárult le, így cikkünkben a működési elv leírása után az egyik nagy súlyú kérdésre, a beszédadatbázis tervezésére koncentrálunk. Ismertetjük a beszédkorpusz tervezésének szempontjait, lépéseit és ellenőrzését.

2. Adott témájú szöveg-beszéd átalakító

Több adott témájú szöveg-beszéd átalakító készült már világnyelveken – például angolul [1] és németül [2]. Ezek beszédadatbázisa két részből tevődik össze. Az

adatbázis nagyobb része a tárgyterület jellemző szavait, kifejezéseit tartalmazza többféle szövegkörnyezetben. A másik része rövid hangsorépítő elemek (diádok, hangok vagy félhangok) teljes halmazát tartalmazza. Ez utóbbiak szolgálnak az olyan szövegrészek összeállítására, amelyek nem szerepelnek a célzott tematikájú adatbázisrészben – jellemzően a témán kívüli szavak, kifejezések.

A szintetikus beszéd előállítása ezekben a rendszerekben úgy történik, hogy az adatbázisból a megfelelő elemeket kiválasztják és összefűzik. Ezek a megoldások az általános TTS-ek körében elterjedt elemkiválasztási algoritmusokat alkalmazzák. Az elemkiválasztási algoritmusokról egy részletes ismertető és a kapcsolódó fogalmak (például költségfüggvények) definíciója [3]-ban olvasható,, illetve egy ilyen elven működő magyar nyelvű rendszer kerül ismertetésre [4]-ben. A bemeneti szövegből az említett rendszerek előállítják a szintézis célsorozatát – a szöveg fonetikus átiratát prozódiai információkkal. Meghatározzák az adatbáziselemek összes olyan sorozatát, ami a célsorozatot teljes egészében lefedi. Ezek közül a jelöltek közül azt választják, amelyik valamilyen költségfüggvényt minimalizál. A jelölt költsége a benne szereplő elemek célköltségeiből (mennyire jól reprezentálja a célsorozat megfelelő szakaszát) és az egymás utáni elemek összefűzési költségéből (mennyiben töri meg az akusztikai jel folytonosságát a két elem összefűzése) adódik össze.

Az előre ismert téma lehetővé teszi, hogy a korpusz nagy része témaspecifikus mondatokból álljon. Így nagy a valószínűsége, hogy a szintézis során hosszú, több szóból álló elemeket választ ki az algoritmus és így sokkal természetesebb hangzást ér el, mintha diádelemekből kellene építkezni. Black és Lenzo rendszere az adatbázisban egymással szomszédos (azaz az eredeti bemondásban is egymás mellett szereplő) elemek összefűzési költségét nullára állítja, így indirekt módon elősegítik a hosszú elemek kiválasztását [1].

Schweitzer és szerzőtársai ezzel szemben a jóval kevésbé számításigényes fonológiai struktúra-egyeztetés (phonological structure matching, PSM) módszert választották [2]. A PSM algoritmus először kideríti, hogy a teljes mondat vagy azok prozódiai egységei[1] teljes egészében megtalálhatóak-e az adatbázisban. Ha igen, akkor ezeket adja ki a kimeneten. Amelyik prozódiai egységet nem találta meg, annak szavait megkeresi és ezeket összefűzve állítja elő a kimenetet. Ha egy szó nem található vagy nem megfelelő környezetben és pozícióban található az adatbázisban, akkor azt szótagokból vagy végső esetben beszédhangokból fűzi össze.

Esetünkben a téma egy infokommunikációs szolgáltató telefonos beszédválaszú rendszereinek üzenetei. Első megközelítésként egy olyan rendszert szeretnénk építeni, amely a promptszövegek egy jól körülhatárolható részhalmazát emberi beszédet megközelítő minőségben képes szintetizálni. Ez a részhalmaz a „gomb" morfémát tartalmazó mondatok köre. A promptszövegek jelentős része tartalmazza ezt a betűsort – a menürendszerben való navigációt szolgáló, vagy a mobiltelefon beállításokat ismertető üzenetekben. Bár a „gomb" mondatok teszik ki a jelenlegi beszédválaszú rendszerek üzeneteinek jelentős részét, ezek entrópiája jóval alacsonyabb a többi mondatnál (amelyek általában hosszabb ismertetők részei, így témájuk és szókincsük jóval változatosabb). Ilyen megfontolások miatt valószínűsíthető, hogy a „gomb"-os mondatokra jó minőségű adott témájú TTS fejleszthető. Ennek a rendszernek a fejlesztése során szerzett tapasztalatok és az elkészült komponensek később felhasználhatók lesznek egy tetszőleges promptszöveget generáló rendszer kidolgozásához.

A „gomb"-os mondatok alacsony változékonyságát figyelembe véve azt tűztük ki célul, hogy a rendszer az esetek többségében képes legyen a kimenetét előre felvett prozódiai egységekből összefűzni. Ha egy szükséges prozódiai egység nem található meg az adatbázisban, akkor azt szavakból fűzzük össze – hasonlóan a PSM megközelítéshez. Ha ez nem sikerül, akkor az adott ponton a rendszer még rövidebb elemekből állítja elő a beszédet. Az elemkiválasztási algoritmus és a szintézis folyamata terveink szerint nagyon hasonló lesz az időjárásjelentés témakörére kidolgozott felolvasóban alkalmazotthoz [4]. A felvételek címkézésére is az ott kidolgozott módszereket fogjuk felhasználni.

3. Korpusztervezés

A fejlesztés első lépése a szöveges adatbázis (szövegkorpusz) megtervezése. Ennek a felolvasott változata a folyamatos beszédet tartalmazó beszédadatbázis, melyből az elemkiválasztó algoritmus tetszőleges hosszúságú (prozódiai egység, szó, diád stb.) elemeket kivághat és beilleszthet a kimeneti jelbe. Az alábbiakban az ehhez kidolgozott módszereket ismertetjük.

A szöveges adatbázis tervezése során két fontos szempontot kell figyelembe venni:

1. Megfelelően kell lefedje a lehetséges, adott témájú szövegeket. Tehát a témához tartozó mondatok szintetizálása során minél hosszabb elemek teljes egészében legyenek benne a szövegkorpusz felolvasása és címkézése után létrehozott beszédadatbázisban.
2. Az általánosság biztosítására legyen lehetőség a témától független szövegek – esetleg rövidebb elemekből történő – összefűzésére is.

Az ismert megoldások [1,2] külön kezelik a két követelményt: felvesznek a témához illeszkedő mondatokat és egy külön mondathalmazzal biztosítják a teljes diádfedést. Úgy gondoljuk, hogy nem feltétlenül szükséges ez a szétválasztás. Egy nagyméretű, adott témájú beszédkorpusz valószínűleg a diádok túlnyomó

1 Prozódiai egységnek nevezzük a beszédnek azt a szakaszát, amely a dallammenet szempontjából egy egységet alkot. A prozódiai egységek gyakran egybeesnek a tagmondatokkal.

többségét tartalmazza. A felolvasási listát ki lehet egészíteni a maradék diádokat tartalmazó mondatokkal.

Ahhoz, hogy minél több, a területre jellemző mondatot tudjunk felvenni, egy infokommunikációs szolgáltató nagy mennyiségű promptszöveget bocsátott a rendelkezésünkre. Ez a szöveghalmaz a 2000. január és 2005. június között felvett és a szolgáltató beszédválaszú rendszereibe beépített promptokat tartalmazta. Ebből a tanítóhalmazból alakítottuk ki a felolvasandó mondathalmaz első változatát.

3.1. Szövegtisztítás és -normalizálás

A további feldolgozás azt igényelte, hogy a rendelkezésünkre bocsátott prompt szöveghalmazból kinyerjük a tényleges promptszövegeket és azokat egységes formátumra hozzuk. Ezt automatikus és félautomatikus módszerekkel végeztük el. Először kiszűrtük azokat a szövegrészeket, amelyek nem részei a promptok szövegének, például a promptok azonosítószáma vagy státusza. Az idegen nyelvű promptokat is eltávolítottuk, de az idegen szavakat tartalmazó magyar üzeneteket meghagytuk, mert ezeket kezelnie kell a rendszernek. Töröltük továbbá a zárójeleket tartalmazó promptokat, mert ezek különleges kezelést igényelnek a felolvasáskor.

A promptlistát egy táblázatként kaptuk meg, cellánként egy prompttal (ami több mondatból is állhat). Ha a cellák szövegének végén nem volt pont, azt pótoltuk. A táblázatot szövegfájllá alakítottuk és automatikusan mondatokra tördeltük, így minden sorba pontosan egy mondat került. Végül az egészet kis betűssé alakítottuk, mert a későbbiekben a kisbetű-nagybetű különbségeket szeretnénk figyelmen kívül hagyni. Erre azért van szükség, mert nyilvánvalóan a „KÜLDÉS", „Küldés" és „küldés" szavak felolvasása teljesen megegyezik.

Így előállt a teljes promptgyűjtemény normalizált formában, amely több, mint 39 ezer mondatot tartalmazott. Ebből kiválogattuk a „gomb" karakterláncot tartalmazó mondatokat és a továbbiakban csak ezekkel dolgoztunk. Ez 4189 mondatot jelent.

3.2. Kategorizálás

A „gomb"-ot tartalmazó mondatok halmazát tovább osztottuk, hogy a különböző, tipikus mondatszerkezetek gyakoriságát és tulajdonságait megvizsgálhassuk. A 4189 mondatot (1596 egyedi) automatikusan az alábbi öt kategória egyikébe soroltuk (*dőlt betűvel* egy-egy példát is megadunk):[2]

– „gomb" szóval végződő mondatok:
  Magánszemélyek, 3-as gomb.
– „nyomja meg a(z) X gombot" kifejezéssel végződő mondatok, ahol X helyén egy szónak megfelelő karaktersorozat áll:
  A kód beírása után nyomja meg a # gombot.
– „nyomja meg a(z) X gombot" kifejezést tartalmazó, de nem azzal végződő mondatok:
  Nyomja meg az OK gombot, és válassza a módosítás opciót.
– „nyomja meg a(z)" kifejezést tartalmazó, de nem az előbbi két kategóriába tartozó mondatok:
  Nyomja meg a boríték jelű gombot hosszan.
– a fenti kategóriák egyikébe se tartozó mondatok:
  *Pluszjelet a * gomb kétszeri megnyomásával írhat.*

Az egyes kategóriákba sorolt mondatok száma az *1. táblázatban* látható. Az „Összes" oszlopban olvasható a megfelelő kategóriájú mondatok összes előfordulásának száma, míg az „Egyedi" oszlopban a legalább egy karakterben különböző mondatok száma.

Kategória	Összes	Egyedi	Lefedő
„gomb végű"	3236	1153	1008
„nyomja meg a(z) X gombot" végű	148	88	81
„nyomja meg a(z) X gombot"	91	36	36
„nyomja meg a(z)"	125	75	67
egyéb	589	244	239
Összesen	**4189**	**1596**	**1431**

1. táblázat
A „gomb" szót tartalmazó és azok teljes lefedéséhez szükséges mondatok száma kategóriánként

Ezután a mondatokat automatikusan prozódiai egységekre tördeltük, így minden prozódiai egység külön sorba került. A tördelés a prozódiai egységek határjelzői (például mondatvégi írásjelek, vessző vagy kettőspont) alapján történt. Mindegyik egységhez eltároltuk a mondatbeli pozícióját is: kezdő/egyedüli, belső vagy záró. Erre azért volt szükség, mert a szintézis során csak a megfelelő pozícióból kivágott prozódiai egységeket célszerű felhasználni a kielégítő prozódia megvalósításának biztosítására. Például nem lehet egy mondatzáró prozódiai egységet egy mondat első egységeként beilleszteni, mert az egész dallammenet természetellenes lesz. A pozícióból eredő különbségeket jól illusztrálja a szóhasználatbeli eltérés – ahogy az a *2. táblázatból* látszik, a leggyakoribb szavak listája jelentős eltérést mutat.

2. táblázat
A promptszövegekben a 10 leggyakrabban előforduló szó a prozódiai egység mondatbeli pozíciója szerint

	Mondatkezdő/egyedüli prozódiai egységekben	Belső prozódiai egységekben	Mondatzáró prozódiai egységekben
1.	a	a	a
2.	az	ft	gomb
3.	és	és	és
4.	gomb	az	az
5.	szolgáltatás	következő	ft
6.	domino	wap	vagy
7.	sms	válasszon	hogy
8.	t	sms	t
9.	kérjük	hogy	sms
10.	ft	t	is

Látható, hogy legalább három prozódiai egység pozíció szerinti megkülönböztetése indokolt. A beszédszintézis területén szerzett tapasztalataink azt mutatják, hogy ennél több pozíció használata nem szükséges.

2 A könnyebb olvashatóság érdekében itt a mondatok a kisbetűssé alakítás előtti formában szerepelnek.

Utolsó lépésként minden kategória prozódiai egységeinek listájából kiszűrtük az ismétlődéseket (a kizárólag a pozícióban eltérőeket különbözőnek tekintve).

Eltárolhattuk volna még a mondatok modalitását is (kijelentő, kérdő, felszólító, felkiáltó vagy óhajtó), erre azonban a promptok témája esetén nincs szükség. Minden promptszövegben előforduló mondat kijelentő vagy felszólító, melyek dallammenete nagyrészt megegyezik.

3.3. A teljes fedést biztosító mondatok kiválasztása

Mind az öt kategóriához adott a szövegkorpusz mondatainak listája és azok felbontása prozódiai egységekre. A cél egy olyan minimális mondathalmaz összeállítása, amely tartalmazza az összes prozódiai egységet legalább egyszer. Ennek a halmaz-fedési problémának egy jól ismert közelítő megoldása a mohó algoritmus [5]. Minden lépésben egy mondatot adunk hozzá a fedőhalmazhoz. Az első kiválasztott mondat az, amelyik a legtöbb új egységet tartalmazza; utána minden lépésben azt a mondatot adjuk hozzá, amelyik a legtöbb, még lefedetlen egységet tartalmazza; ezt addig ismételjük, amíg van lefedetlen egység. A mohó algoritmus többféle továbbfejlesztése ismert [6], de ezek ennél a problémánál nem jelentenek előnyt.

A mohó algoritmus futtatásával megkaptuk a felolvasandó mondatok listájának első változatát. Kategóriánként a kiválasztott mondatok számát az 1. táblázat foglalta össze. Az eredeti mondatszámhoz képest kis mértékű, körülbelül 10%-os csökkenést értünk el.

A módszerrel kapcsolatban két dolgot érdemes kiemelni. Az egyik az, hogy végig a szövegek írott formájával dolgoztunk, nem a fonetikus átiratukkal. Ez bizonyos mértékű hibát okoz az eredményekben – ha egy prozódiai egységnek két, különbözőképpen leírt formája szerepelt a szövegekben, akkor az az egység duplán szerepel a felolvasási listán. Ezek száma valószínűleg kevés. Az előállított beszéd minőségét ezek a „hibák" nem rontják, sőt, egyes esetekben javíthatják (ugyanabból az elemből több alternatíva áll rendelkezésre, így az elemkiválasztó algoritmus az adott célsorozathoz leginkább illeszkedőt tudja kiválasztani).

Ha azonban betű-beszédhang átalakítást alkalmaztunk volna a mohó algoritmus futtatása előtt, az számos problémát vetett volna fel a betű-beszédhang algoritmusok tökéletlensége miatt. Schweitzer és szerzőtársai 170 ezer német mondatból 70 ezerben találtak átírási hibát [2]. Magyar nyelv esetén ez kisebb problémát okoz, mert maga a betű-beszédhang átalakítás egyszerűbb. Viszont elengedhetetlen az ezt megelőző betű-betű átalakítás, amely többek között a rövidítések, számok kifejtését, az idegen és rendhagyó írásmódú szavak átírását jelenti. Ez az algoritmus számos esetben téved és ezeket a tévedéseket nehéz automatikusan felderíteni.

A másik megjegyzés a gyakoriságokkal kapcsolatos. Nyilván érdemes lenne figyelembe venni az egyes prozódiai egységek előfordulásának gyakoriságát is. Viszont az, hogy a promptok szövegeiben hányszor fordul elő egy-egy egység, független attól, hogy milyen sűrűn játssza le azt az egységet a rendszer. Tehát a rendelkezésre álló adatok alapján számolt gyakoriságok valószínűleg félrevezetőek. Van Santen és Buchsbaum szerint még akkor sem érdemes a gyakorisági információt felhasználni, ha azok elérhetőek [5]. Ennek oka, hogy ezek a gyakoriságok csupán egy időben változó eloszlás mintái – jó példa erre a „WAP" és a „telex" szavak, melyek előfordulásainak száma nagyságrendileg eltérő lehet egy 1995-ös és egy 2005-ös promptgyűjtemény között.

4. A szövegkorpusz ellenőrzése

Annak érdekében, hogy a koncepció helyességét ellenőrizzük, egy független teszthalmaz fedését is kiszámítottuk. A teszthalmaz egy infokommunikációs szolgáltató két telefonos beszédválaszú rendszerében 2005. novemberben használt összes prompt szöveges formája. A tanító- és a teszthalmaz méreteit a 3. táblázat foglalja össze.

	Tanítóhalmaz	Teszthalmaz
Mondatok (ismétlődésekkel)	39 247	5 666
Szavak	530 526	61 225
Szóalakok	8 914	5 621
Prozódiai egységek	82 642	11 968
Egyedi prozódiai egységek	17 849	7 286

3. táblázat A tanító- és a teszthalmaz méretei

Kiszámítottuk, hogy a teszthalmaz prozódiai egységeinek hány százalékát fedik le a felolvasólista mondatai. Akkor tekintettünk egy egységet lefedettnek, ha a felolvasólistán van egy olyan egység, amelyik ugyanazt a karaktersorozatot tartalmazza és a mondatban ugyanabban a pozícióban van. Az eredmények a prozódiai egység kategóriák és pozíciók szerint a *4. és 5. táblázatban* láthatóak.

A teljes fedési arány nagyon magas, 76%. Ez azt jelenti, hogy ha a promptgenerátort a felolvasási lista első változata alapján valósítanánk meg és a teszthalmaz teljes szövegét szintetizálnánk a rendszerrel, akkor négyből három prozódiai egység emberi beszédet megközelítő minőségben szólna.

4. és 5. táblázat
A teszthalmaz összes egyedi prozódiai egységeinek és a lefedetlenek száma kategóriák és pozició szerint

Kategória	Összes	Lefedetlen
„gomb végű"	839	206 (25%)
„nyomja meg a(z) X gombot" végű	63	17 (27%)
„nyomja meg a(z) X gombot"	50	14 (28%)
„nyomja meg a(z)"	35	4 (11%)
egyéb	235	51 (22%)
Összesen	1222	294 (24%)

Pozíció	Összes	Lefedetlen
Mondatkezdő/egyedüli	785	132 (17%)
Belső	288	119 (41%)
Mondatzáró	149	43 (29%)
Összesen	1222	294 (24%)

Az egyedi szóalakok fedése 95%. Előzetes feltételezésünk az volt, hogy ha magas fedési arányt tudunk elérni a prozódiai egységek szintjén, akkor az implicit módon magas szószintű fedési arányt is jelent. Ez a feltételezés bebizonyosodott. Bár a szószintű fedésnél csak a szavak karaktereinek egyezését néztük (így például a pozíciót, környezetet, hangsúlyosságot figyelmen kívül hagyva), valószínűsíthető, hogy azokban az esetekben, amikor a cél prozódiai egység nem található meg az adatbázisban, a rendszer legtöbbször képes azt szavakból összefűzni. Ezek alapján feltételezzük, hogy a promptgenerátorral a jelenleg használatos diád- és triád-alapú rendszerek beszédének természetességét messze felül tudjuk múlni.

Érdekes megfigyelés, hogy a lefedetlen szóalakok száma (310) nagyjából megegyezik a lefedetlen prozódiai egységek számával. Lehetséges, hogy a legtöbb lefedetlen prozódiai egység mindössze egyetlen szóban tér el a tanítóhalmazban hozzá legközelebb levő prozódiai egységtől és ez az egy szó egyáltalán nem szerepelt a tanítóhalmazban.

5. Összefoglalás és kitekintés

Cikkünkben bemutattuk egy készülő promptgenerátor beszédkorpuszának tervezési folyamatát. A rendelkezésre álló promptszövegeket normalizáltuk, kategóriákra bontottuk, majd a mondatokból mohó algoritmussal egy (prozódiai egységek szempontjából) teljes fedést biztosító részhalmazt választottunk ki. Az összeállított mondathalmaz egy független teszthalmaz prozódiai egységeinek 76%-át lefedte. A szószintű fedés 95% volt. Ezek az adatok azt mutatják, hogy elképzelésünk működőképes.

A végleges felolvasólista a 3. szakaszban ismertetettnél nagyobb lesz, mert azon felül az alábbiakat is tartalmazni fogja:

- Olyan mondatok, amelyek hozzáadásával elérhető, hogy a teszthalmazra is 100%-os legyen a prozódiai egységek fedése. Ez körülbelül 300 új mondat hozzáadását jelenti.
- Az esetlegesen hiányzó diádokat tartalmazó bemondások.
- A számok megfelelő kezeléséhez a teljes számelemkészlet [7].
- A betűszavak jó minőségű szintéziséhez az összes betűelem (például „emm", „zé" és „iksz").

A bemutatott módszernek több ismert hiányossága van. A mohó algoritmus a hosszú mondatokat preferálja a rövidek helyett, pedig az utóbbiakat sokkal könnyebb helyesen felolvasni a korpusz felvételekor. Ez elsősorban a keresés elején igaz, amikor még szinte minden prozódiai egység lefedetlen, így egy hosszabb mondat több újonnan lefedett egységet eredményez. Az azonos számú lefedetlen egységet tartalmazó mondatok közül pedig a listában először szereplőt választja. Az utóbbi probléma azonban orvosolható: azonos számú új egységet tartalmazó mondatok közül mindig a rövidebbet kell választani.

Jelenleg a mohó algoritmust mind az öt kategóriára külön futtattuk, hogy jobban megismerjük azok eloszlását. A végleges felolvasási lista mérete azonban kissé csökkenthető, ha egyszerre futtatjuk az összes kategóriára az algoritmust, ugyanis az egyes kategóriák között lehetnek átfedések.

Az egyes prozódiai egységek lejátszásának gyakorisága alapján tovább javítható a korpusz összetétele: például a leggyakoribb egységeket többször is felvehetjük, míg a nagyon ritkákat kihagyhatjuk. Jelenleg ilyen statisztikai adatok nem állnak rendelkezésünkre, azonban a rendszer béta változatának beindításától kezdve a naplófájlokból kinyerhetőek.

A továbblépés szempontjából fontos kérdés, hogy az összes prompt mondat (nem csak a „gomb"-ot tartalmazók) teljes fedéséhez hány további mondat felvétele szükséges. Egyenlőre nem egyértelmű, hogy a prozódiai egység szintű megközelítés ebben az általánosabb esetben milyen eredményre vezet.

Köszönetnyilvánítás

Ezt a kutatást az NKFP 2. program (szerződésszám: 2/034/2004) támogatta.

Irodalom

[1] Black, A. W., Lenzo, K. A.,
Limited domain synthesis, Proc. of ICSLP 2000.
[2] Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., Sauberlich, B.:
Restricted unlimited domain synthesis,
Proc. of Eurospeech 2003, pp.1321–1324.
[3] Möbius, B.,
Corpus-based speech synthesis: Methods & challenges,
Arbeitspapiere des Instituts für
Maschinelle Sprachverarbeitung (Univ. of Stuttgart),
Vol. 6, No.4, 2000, pp.87–116.
[4] Fék, M., Pesti, P., Németh, G., Zainkó Cs.:
Generációváltás a beszédszintézisben,
Híradástechnika, jelen számban
[5] van Santen, J. P. H., Buchsbaum, A. L.,
Methods for optimal text selection,
Proc. of Eurospeech 1997, Vol. 2., pp.553–556.
[6] Bozkurt, B., Ozturk, O., Dutoit, T.,
Text design for TTS speech corpus building using a modified greedy selection,
Proc. of Eurospeech 2003, pp.277–280.
[7] Olaszy G., Németh G.:
IVR for banking and residential telephone subscribers using stored messages combined with a new number-to-speech synthesis method,
In Gardner-Bonneau, D. (ed.): Human Factors and Interactive Voice Response Systems, Kluwer, 1999.
pp.237–255.

A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései

OLASZY GÁBOR

BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Lektorált

Kulcsszavak: korpusz alapú beszédszintézis, hangszerkezetek, hangátmenetek, optimális vágási pontok, prozódiai invariancia

A beszédprodukció eredménye a beszéd hullámformája. Ez mindenkor egyedi és egyszeri produktum. A beszédszintézis által előállított hullámforma és az emberi produktum közötti az alapvető ellentmondás abban van, hogy a szintézisnél egy tárolt (fix) adatbázisból építjük fel a beszédjelet, tehát megsértjük az egyedi-egyszeri produkcióra vonatkozó tételt. A kérdés az, hogy hogyan lehet ezt az ellentmondást csökkenteni. A korpusz alapú szintézis elvéből fakad, hogy az egyszeri jelzőre vonatkozó időtényezőt próbálja tágítani azzal, hogy hosszabb elemekből építkezik mint a korábbi szintetizálási technológiák, noha itt is egy előre meghatározott, tárolt beszédadatbázisból (korpusz) történik a szintetizált beszéd előállítása. Ennek a törekvésnek a támogatására foglaltuk össze azokat a legalapvetőbb nyelvi, fonetikai ismereteket, amelyekkel segíteni lehet a legjobb jelöltek megtalálását a korpuszban és ezzel a minél jobb hangminőség elérését.

1. Bevezetés

Egy szöveg és annak felolvasott, hangzó formája között szoros összefüggés van. A szöveg a közölt gondolat tartalmát hordozza, valamint tartalmazza a mondat modalitására (kérdés, kijelentés), tagolására utaló írásjeleket is. A szöveg akusztikai szintre való transzformálásakor létrejött hangsorozat képviseli a beszéd úgynevezett szegmentális szerkezetét, (ezt tekinthetjük az írott szöveg tartalmi lényegének, legfőbb hangzó megtestesítőjének). Az előbbi fogalomhoz egy zenei példát adva a kottafejek puszta pontos lejátszása adja a zene szegmentális akusztikai megvalósítását. A valódi értelmezést a művész egyéni megformálása hozza létre az egyéni játékával és a hangszer tulajdonságainak felhasználásával. A beszédben is ez a helyzet, a beszéd „hangszerelését” a hangsúlyozás, a dallammal való variálás, a ritmus megvalósítása és a hangszínezet adja. Ezeket nevezik szupraszegmentális eszközöknek, egy szóval prozódiának.

A beszélőnek a szövegben egyrészről írásjelek jelzik a modalitást, a tagolást, ezek alapján hozza létre a fő dallamszerkezetet, a ritmikát (beszédsebesség, szünettartás, szünethossz). Másrészről a beszélő egy azonnali értelmezést is végrehajt a felolvasásnál. Ez is befolyásolja az előbbi szupraszegmentális elemek variálását. A fenti két nyelvi, fonetikai tényezőt tehát a beszélő automatikusan beleépíti a produkciójába a szegmentális és szupraszegmentális szint egyszerre jelenik meg a jelben. A beszédszintézisben géppel helyettesítjük a beszélő személyt és itt is arra kell törekednünk, hogy az előbbi két nyelvi, fonetikai tényező is bekerüljön az előállított beszédjelbe. A hallgató elvárja, hogy a géppel összeállított beszéd hangzása minél közelebb álljon az emberi hangszínezethez, továbbá, hogy a hangsúlyozás, a ritmikai szerkezet és a beszéddallam kialakítása megfeleljen az adott témának, az ahhoz köthető stílusnak. A beszédszintetizátor produkcióját a hallgató a hangzás alapján értékeli és nem érdekli, hogy a részletek mögött milyen technológia áll. Ezért a legapróbb akusztikai, fonetika részleteket is gondosan tervezni kell, hogy a szintetizált beszédbe minél kevesebb akusztikai hiba kerüljön bele. Ez azt jelenti, hogy nem csupán az akusztikai jellel, mint rezgésformával kell foglalkoznunk a szintézis során, hanem annak nyelvi, fonetikai hátterével is. Ebben a cikkben erre próbálunk rámutatni.

2. A korpusz alapú beszédszintézis fonetika kérdései

Napjainkban a korpusz alapú beszédgenerálás adja a legjobb minőségű szintetizált beszédet.

A technológia nevéből adódik, hogy egy adott beszédkorpusz képezi a szintézis alapját [1] (nagy korpusz, sok órányi beszéddel, annak szöveges és hullámforma-szintű tárolásával és címkézésével). A mérnöki gondolat az így szintetizálandó mondat előállítására az, hogy a korpuszban lépésről lépésre megkeresik az adott szintetizálandó beszédrészleteknek (mondatrész, szófüzér, szó, hangkapcsolat) legjobban megfelelő hullámforma részeket (ezek hossza az esetek többségében lényegesen nagyobb kell hogy legyen, mint egy hangkapcsolat), és azokat kivágva, majd egymás után kapcsolva létrehozzák a mondatot reprezentáló szintetizált hullámformát. Az alapgondolat szerint a cél annak elérése, hogy ne kelljen semmiféle mesterséges jelfeldolgozást végezni a hullámformán, csupán válogatással és összefűzéssel lehessen megoldani az új mondat összeállítását. Az elvárás az, hogy jó válogatás esetén az így előállított beszéd minősége igen közel lesz a korpusz eredeti beszédének a minőségéhez. Ez az új technológia annak a hiányosságnak a ki-

küszöbölésére született meg, hogy az eddigi beszédszintetizáló rendszerek hangminősége ugyan már közel állt az emberéhez, érthető is volt, de közel sem volt olyan hangzású, mint egy ember által felolvasott szöveg, nem lehetett például egyértelműen személyhez kötni, hiányzott belőle a beszélő személy egyéni hangszínezete. A fenti mérnöki gondolatot támogatta még a számítástechnika rohamos fejlődése memóriakapacitásban és feldolgozási sebességben.

A kérdés az, hogy milyen fonetikai, nyelvi szempontokkal kell számolni egy ilyen rendszer tervezésénél és megvalósításánál. A legfontosabb az, hogy figyelembe kell venni a beszédképzés biológiai természetéből adódó tényt, mármint, hogy **az emberi beszéd egyedi és egyszeri produktum**. Az egyedi jelző az egyén saját hangjára vonatkozik, az egyszeri pedig azt fejezi ki, hogy a beszédjel az adott időpillanatra jellemző akusztikai szerkezettel valósul meg. Ez a pillanat a kiejtés pillanata (az akkori biológiai jellemzők határozzák meg a hangot: a gége állapota, a beszélő nyelvi döntései az értelmezést illetően, az artikulációs szervek és környezetük állapota stb.) Még ugyanazon beszélő sem tud egy mondatot ugyanolyan akusztikai szerkezettel kimondani, felolvasni, más lesz a ritmikája, más lesz a hangszínezete, más dallamot valósíthat meg. Az artikuláció tehát pillanatnyi, ugyanakkor folyamatos. Ebből következik, hogy a létrehozott beszédjel is az.

Az említett kiejtési változatosság ellentétben áll azzal a ténnyel, hogy a mai technológiákban a szintézishez a beszédjelet egy rögzített, korpuszból, adott számú tárolt hullámformából, jelválogatással és a jelek összekapcsolásával állítják elő. A rögzített korpusz tartalma mindig ugyanaz, tehát minden esetben ugyanazon korlátozott számú elemhalmaz áll a szintézis rendelkezésére. Az így összeválogatott hullámforma elemek ugyan hangilag tartalmazni fogják a szövegnek megfelelő hangsort, de a beszédjel akusztikai részleteiben a folyamatosság nem jön létre. Olyan eltérések lehetnek az összeillesztési pontoknál (a különböző helyekről való egyedi kivágások következményeként), amelyek megtörik az egyenletes, folyamatos hangzást, ezért zavaróak lesznek a hallgató számára (a beszéd nem lesz természetes hangzású, dallamugrások, hangszínezet-váltások, hangerőváltozások lesznek benne a vágási pontokhoz kötve). Úgy gondolták, hogy a rögzített korpusz nagyságának növelésével ezek a gondok csökkenthetők. Jó példa erre egy Japánban készített korpusz alapú szintetizáló rendszer, amelyik 380 órás hangfelvételű beszédkorpuszból építi fel a beszédjelet [2]. A korpusz növelése sem egyszerű feladat. Számolni kell a beszélő hangszínváltozásával a hosszú felvétel alatt, továbbá a feldolgozási időtényezővel, ami mind az adatbázis elkészítését (egy profi bemondó például 3-4 óránál többet nem tud egy nap teljesíteni, már ez idő alatt is a hangszínezete lényegesen megváltozhat), annak további feldolgozását, előkészítését, mind pedig a későbbi szintézisnél alkalmazott válogatási algoritmust érinti.

3. A vágási pontok teóriája

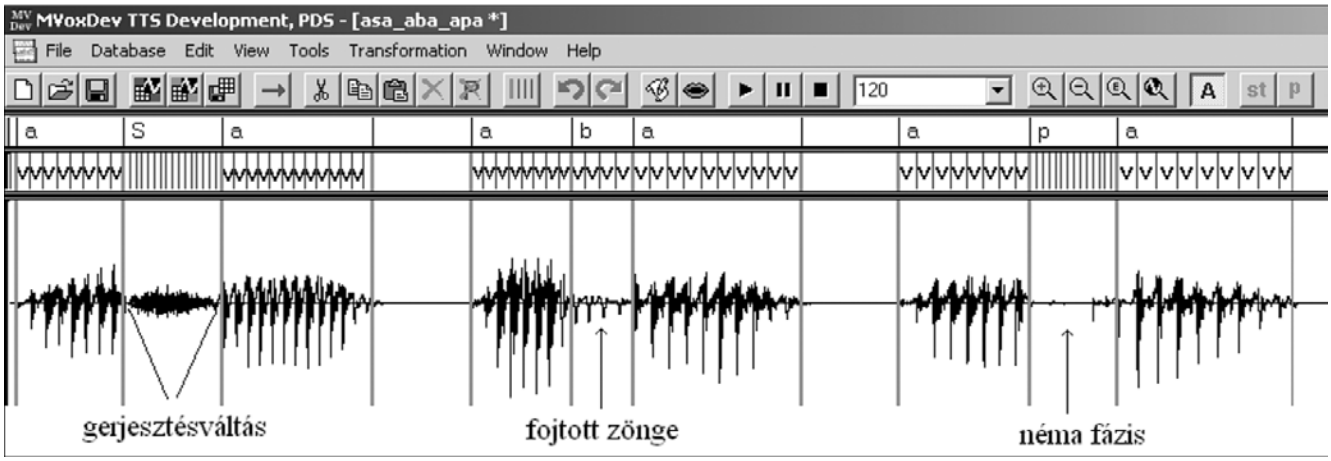
A mai emberi hangú beszédszintetizátorok többsége előre eltárolt, rövid hullámforma-részletekből építkezik, azaz két félhangot (diádot) kapcsolnak össze a hangsor felépítésénél, esetleg kicsivel hosszabb elemet, a két félhang között egy egész hangot tartalmazót, azaz triádokat. A fő kritika a jelenlegi, ilyen elemösszefűzéses technológiákkal szemben az hogy az építőelemek rövidek így az összeillesztési pontok száma az egységnyi (például 1 s-ra számított) beszédjelben sok, ezért sok torzítás kerül a jelbe (minden vágásnál a két összevágott elem határán spektrális eltérések lehetnek, ezek torzítást okoznak). A vágási pontok csökkentésére való törekvés eredménye a korpusz alapú szintézis elve. Ez a teória azonban csak felszíni jelenségre, a vágási pontok számának csökkentésére alapozza megállapítását.

A torzulások az összevágott beszédjelben nem a vágások nagy száma miatt jönnek létre, hanem azért, mert az összevágott jelrészleteket különböző időpontokban ejtett beszédjelből származtattuk, tehát megsértettük azt az előbbi tételt, hogy a beszédjel egyedi és egyszeri. Ha például egy mondat hullámformáját minden hang kezdeténél való szétvágással szétszereljük (sok vágási pontot generálunk), majd újra összeillesztjük, a mondat hangzásában semmilyen minőségromlás nem lesz. Tehát a sok összeillesztési pont önmagában nem kell, hogy generálja a torzítást.

A korpuszos szintézisnél létrejövő minőségjavulás azért jön mégis létre, mert nem rövid jelrészleteket illesztünk egymás után (diádos, triádos hangkapcsolatot), hanem hosszabb beszédszakaszokat (szó, szókapcsolat, szófűzér stb.), tehát ezzel kevésbé sértjük meg a beszédjel egyszeri megvalósulására vonatkozó állapotot. Ha például szavakból, szókapcsolatokból illesztünk össze egy mondatot, akkor az egyszeri, ugyanazon biológiai paraméterekkel megvalósuló beszédprodukciót átfogó időtartam növekszik, tehát hosszabb ideig fogja automatikusan hordozni a beszélőre jellemző hangszínezetet. Ha ehhez még hozzá vesszük, hogy a hallgató hosszabb nyelvi egységekben is kapja ezt a minőséget (szó, szófüzér hosszúságú elemekben), akkor be kell látnunk, hogy a percepció számára kellemesebb, jobb minőségű beszéd jöhet létre, mint amilyen a diádos, triádos rendszereké. Tehát nem az öszszeillesztési pontok számának csökkenése adja a kedvezőbb eredményt, az csak következménye a hangsorépítésnek.

4. Az optimális fizikai vágási pontok meghatározása

A cél, hogy olyan ponton vágjuk el a hullámformát (vegyük ki eredeti korpuszos környezetéből), amely a legkevesebb torzítással jár a későbbi összeillesztésnél. A döntést két tényező befolyásolja: milyen a hangszintű szerkezet az adott ponton és, hogy milyen a prozódiai szerkezet. Itt most a hangszintű szerkezetről fo-



1. ábra Az optimális vágási pontok bemutatása az asa, aba, apa *hármas hangkapcsolatokban. A függőleges vonalak hanghatárok, a zöngés hangok periódushatárait a* v *jelzésű vonalmarkerek jelzik a hullámforma felett.*

gunk elsősorban beszélni. A fizikai vágási pont kialakításához ismerni kell a beszédhangok artikulációs és spektrális belső szerkezetét, valamint tisztában kell lenni a hangkapcsolódások megvalósulásakor létrejövő hangszerkezeti és spektrális módosulások fajtáival. A vágást akkor végezhetjük sikeresen, ha tudjuk, hogy a beszédhangoknak milyen az egymásra hatása, a belső akusztikai szerkezete, hol milyen változás zajlik le a hang frekvencia-, illetve intenzitásszerkezetében a folyamatos artikuláció következtében, melyek azok a hangrészek, amelyek esetleg egymással megegyeznek, illetve nagyon hasonlóak egymáshoz.

A lényeg az, hogy a korpuszból kivágott elemek összeillesztésénél az elemek határán lévő beszédhangokra hangkapcsolódási illesztést kell végrehajtanunk. Úgy kell kiválasztani a kivágandó elemet, hogy ne sértsük meg a spektrális folyamatosság elvét. A következőkben megadjuk a hangsor azon fizikai helyeit, ahol a legoptimálisabbak a vágási pontok: a hangsor minden olyan pontja, ahol gerjesztés váltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitás minimum keletkezik), továbbá a hangok belsejében lévő néma fázisok, zönge szakaszok (ez a zár- és zár-rés hangok sajátja). Az optimális vágási pont kijelöléséhez 5-10 ms pontosságú helymeghatározásra, általában zöngeszinkron jelölésre van szükség *(1. ábra)*. A hangsor összeállításánál ezek után a kivágott beszédrészletek egymáshoz való illesztését hangszerkezeti és artikulációs fonetikai szabályok alkalmazásával tehetjük optimálissá, torzításmentessé. Az illesztés akkor lesz sikeres, ha a beszédjelen nincs hallható akusztikai torzulás a beavatkozás után [3].

Mindezek alapján összeállítottuk a vágási pontokat meghatározó fonetikai szabályokat *(1. táblázat)*. A vágás sikeres elvégzésének alapfeltétele, hogy a hanghatárokat előzőleg pontosan jelöljük be a hullámformán. Az alábbi szabályok minden esetben a hanghatárra, mint vágási pontra vonatkoznak. A vágási pont zöngés hangok esetében vagy a hanghatár, vagy a hang belsejében vonalmarkerrel jelölt periódus határ.

A táblázatból látható, hogy a vágási pont keresését hangszinten kell végrehajtani, ebből adódik, hogy a lexikai formát át kell alakítani fonemikus, hangreprezentációs formává, hogy a szabályok alkalmazhatók legyenek. Ez is fonetikai jellegű tudást igényel.

1. táblázat Példa szabályokra a vágási pontok kijelöléséhez a korpuszban. A hangcsoportok a csatlakozó második hang jelölésére: C= mássalhangzók; V= magánhangzók; C1= p,t,k,ty,h,f, s,sz,c,cs; C2=b,d,g,gy,zs,z,dz,dzs; C3= v, j, l, r ; C4= m, n, ny; kiv:= kivéve

megelőző hang a betűjele szerint	következő, kapcsolódó hang a betűjele szerint	vágási pont kijelölésének szabálya	szöveges példa (a csatlakozó hangokat kiemeléssel jelöltük)
a) b, d, g, gy b) b, d, g, gy	a) V, C3, C4 b) önmagával csatlakozik	a) a hanghatáránál kell vágni b) a hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne)	*vad vihar* *nagy meleg* *vad dörrenés*
n	a) V, C kiv. C4 b) önmagával	a) a hang határánál kell vágni b) a hang 70%-ánál kell elvágni	*télen derült* *télen nagyon*
f, sz, s, c, cs, h	a) V, C3, C4, C1 kiv. b), b) önmagával	a) a hang határánál kell vágni b) a hang 70%-ánál kell elvágni	*havas lesz* *havas síkos.*
l	a) V b) C kiv. c) c) önmagával	a) nem célszerű elvágni b) a hang határnál kell elvágni c) a hang 70%-ánál lehet elvágni	*szél óránként* *tél marad* *fel lesz*

5. Az artikuláció akusztikus vetülete

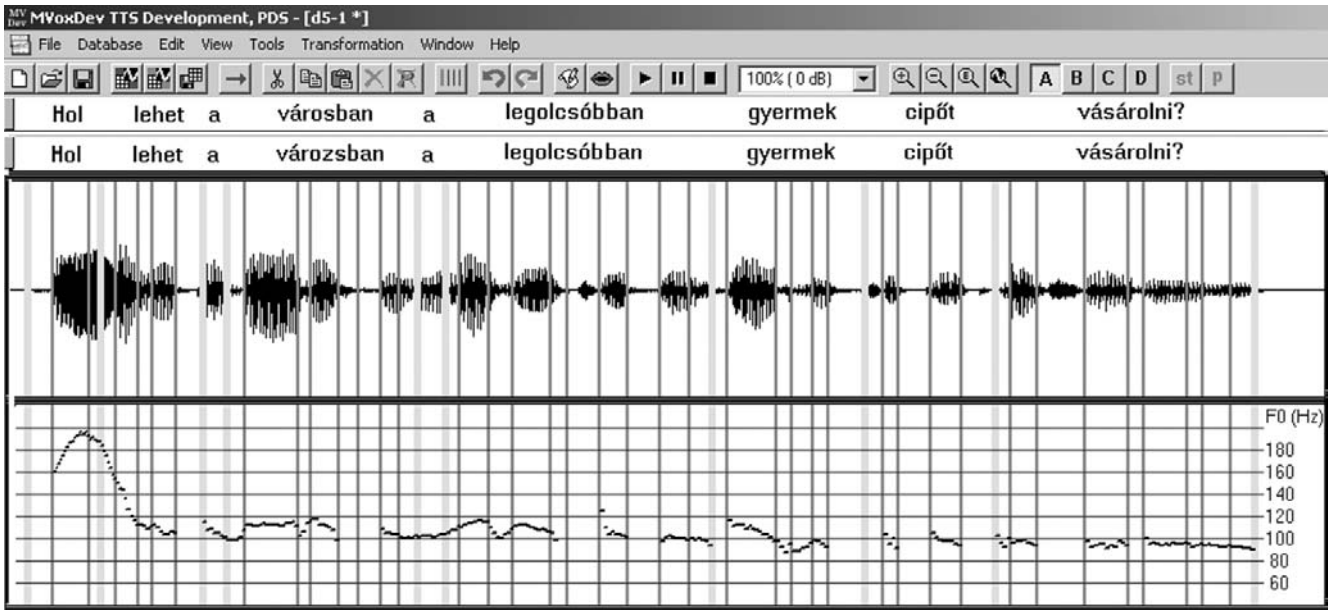
Amennyiben a táblázat szabályai szerinti hangkombinációt a korpusz nem tartalmaz, akkor a torzításmentes vágási pont megtalálásához másodlagos jelölteket is lehet állítani. A megoldás elméleti hátterét a hangok képzésekor kialakuló artikulációs konfiguráció ismerete adja [3]. A képzési hely önmagában reprezentál egy elméleti akusztikai tartalmat. Minden artikulációs pozíciónak megvan a saját statikus spektrális megfelelője az akusztikai térben, amit a formánsok, illetve zörejgócok frekvenciáival fejeznek ki. A folyamatos beszédben a képzési konfigurációk (mozgássorozatok) követik egymást, a mozgássorozatok egymásra hatását pedig akusztikai szinten a hangok úgynevezett átmeneti fázisai tartalmazzák (a formánsok mozognak). A másodlagos jelöltként felhasználható vágási szabályokat CV és VC elemekre vonatkoztatjuk. A cél az egymáshoz hasonló, így egymással helyettesíthető hangkapcsolati elemek meghatározása. A helyettesítő szabályok kidolgozásánál a mássalhangzók képzési helyeiből és azok akusztikai tartalmából, mint statikus tényezőkből indulunk ki, majd a CV, VC összekapcsolódásból adódó dinamikus változásokat tanulmányozzuk, vagyis azt, hogy a mássalhangzók hogyan hatnak a magánhangzók spektrális szerkezetére.

Akusztikai szempontból kitüntetett szerep jut itt az azonos, illetve közel azonos képzési helyű mássalhangzóknak, mivel azok közel azonos statikus akusztikai hatással rendelkeznek, ezért azonos hatást fejtenek ki a magánhangzóra, annak a mássalhangzóhoz csatolódó részére, az átmeneti fázisára (például az *ások, ácsok* szavakban az (o)-ban szinte ugyanaz a spektrális szerkezet alakul ki a (s), illetve a (cs) hatására). Másrészről ezen mássalhangzók hangzó része is egymáshoz hasonló spektrális szerkezetű, ha a gerjesztést nem változtatjuk meg (például az ások, ácsok szavakban az (s), illetve a (cs) hang spektrális komponensei igen hasonlóak). A hangkapcsolódásra jellemző spektrális szerkezet tehát előre ismerhető (a vágási pont optimalitási szintje jósolható). A magyar mássalhangzók képzési hely szerinti csoportosítását a *2. ábra* mutatja. Az azonos képzési helyű hangok egy-egy sort képviselnek, ezek a sorok képezik a másodlagos vágási pontok kijelölésének az alapját. Az előbbi okfejtés szerint tehát az ábra minden vízszintes sora egy-egy artikulációs pozíciót reprezentál. Ennek a pozíciónak az akusztikai vetülete elméletileg hangtól függetlenül átkerül a mássalhangzót követő, illetve megelőző magánhangzóba. Ez azt jelenti, hogy az azonos sorban szereplő mássalhangzók hatása a magánhangzók formánsszerkezetére jó közelítésben azonos. Mivel a nazális mássalhangzók többnyire nazalizálják a magánhangzót is, ezeket nem célszerű a többi ugyanazonsori mássalhangzóval egyforma kategóriába sorolni.

A 2. ábra első és második sorában tehát gyakorlatilag csak a (b, p), illetve a (v, f) hangpár használható vágási szabályhoz. A harmadik sorban hét hang, a negyedikben hat található. Tehát ez a két artikulációs pozíció viszonylag széles körben használható a helyettesítésre. Az ötödik sorban szereplő palatális mássalhangzók pedig képzési helyük kitüntetett volta miatt használhatók jó hatásfokkal (a palatális mássalhangzók erős és jellegzetes hatást gyakorolnak a magánhangzók formánsszerkezetére [4,5]). A hatodik sorban látható (g, k) mássalhangzókat pedig alkalmazkodó képességük miatt lehet hatásosan felhasználni. Ezek a mássalhangzók ugyanis a legképlékenyebbek, a képzési helyük változik a hozzájuk kapcsolódó magánhangzó függvényében. Lássunk a fentiek alkalmazására egy példát. Ha a szöveg szerint *szá* kapcsolat hanghatárán kell illeszteni (például az *sz* az előző szó utolsó eleme az *á* pedig a hozzá csatolandó szó első hangja), de nincsen a korpuszban csak *szo*, viszont van *cá*, vagy *tá*, akkor az utóbbi kettő felhasználásával kiválasztható a helyettesítő elem. A 2. ábra szerint az (sz) hangot helyettesítheti a (c), illetve a (t) hang is, így a helyettesítő kapcsolódó elem lehet a *tá* vagy *cá* kapcsolatból kiválasztott (á) hang is. Az így összeillesztett (sz)+(á) hullámforma nem tartalmaz spektrális torzítást, mivel a fenti mássalhangzók a magánhangzóban ugyanazt a spektrális átmeneti fázist hozzák létre.

2. ábra A magyar mássalhangzók képzési hely szerinti csoportosítása

	zárhangok								zár-réshangok				réshangok							nazálisok					
	b	p	d	t	gy	ty	g	k	c	dz	cs	dzs	v	f	z	sz	zs	s	h	m	n	ny	j	l	r
két ajak	☒	☒																		☒					
ajak-fog													☒	☒											
fog-fogmeder			☒	☒					☒	☒					☒	☒					☒				
fogmeder											☒	☒					☒	☒						☒	☒
elülső szájpadlás					☒	☒																☒	☒		
hátsó szájpadlás							☒	☒																	
gége																			☒						



3. ábra Egy kérdő mondat Fo menetének vizsgálata a szöveg függvényében. A hanghatárokat vékony, a szóhatárokat vastag függőleges vonalak jelzik.

6. A prozódiai tartalom és a szöveg összefüggései

A vágási pont optimális meghatározását – mint korábban említettük – két tényező befolyásolja: milyen a hangszintű szerkezet az adott ponton és hogy milyen az alapfrekvencia (Fo=hangmagasság) és az intenzitás (I) a kapcsolati ponton (a prozódia két fő eleme). A kérdés azért fontos, mert a hangsorban nem csak a hangfolyam képez egy folyamatos egységet, de a prozódiai szerkezet elemei is. A prozódiai szerkezet lefolyása a mondat elejétől a végéig folyamatos és jellemző a mondatra. A prozódiával tehát külön is kell foglalkoznunk, hogy az összekapcsolt hullámforma részletek ilyen szempontból is illeszkedjenek egymáshoz. Hogyan határozhatunk meg olyan vágási pontokat, amelyek a hangszíntű optimalitáson túl a prozódiai folytonosságot is biztosítják? Hogyan jósolható a prozódia a szövegből?

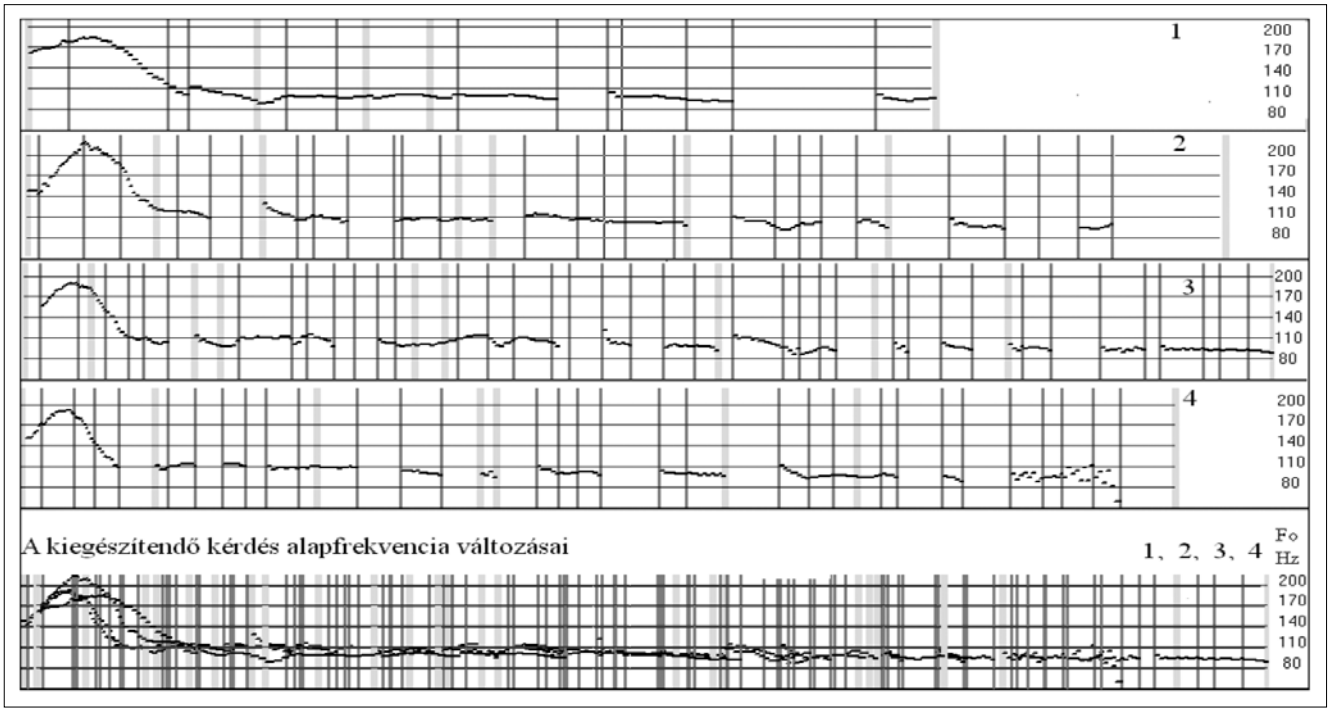
A problémakört két irányból közelíthetjük meg. Az egyik, amikor a szöveg oldaláról végzünk nyelvészeti elemzést, és a szövegben megjelöljük a várható fő dallamformákat, kijelöljük a hangsúlyokat, meghatározzuk a szünetek helyét, és azt mondjuk, hogy a felolvasó személy nagy valószínűséggel ezek szerint fogja felolvasni a szöveget (vö. [6]). Az elemzés eredményét rávetítve a felolvasott beszédjel Fo görbéjére, intenzitás függvényére megállapíthatjuk, hogy vannak-e eltérések a nyelvészeti elemzés és a produkció között. Ha nincsenek jelentős eltérések, akkor a nyelvi elemzés segítheti a helyes prozódiai szerkezetek azonosítását a beszédkorpuszban a szöveg-hullámforma megfeleltetésnél.

A másik megközelítés fonetikai jellegű. Itt a beszédprodukció oldaláról indulunk ki és a fizikai paraméterek változását vetítjük rá a kiindulási szövegre, majd a megjelölt változási pontok alapján vonunk le következtetéseket a szöveg és a prozódia kapcsolatáról. A következőkben erre mutatunk példát. A vizsgálatunk középpontjában a beszédjel prozódiai szerkezete áll, és az, hogy ennek a szerkezetnek a változási pontjait rávetítsük a szövegtestre. A kérdés tömören az, hogy mennyire invariánsak a magyar mondat dallamformái és intenzitás függvényei. Ha sikerül a változási pontokat mind a szövegtestben, mind a prozódiai elemekben egymással szövegfüggetlenül összerendelni, akkor az adott szövegrész megváltoztatásával és az arra vonatkozó prozódiai szerkezet megtartásával elérhetjük, hogy a beszéd tartalmi része megváltoztatható (más szövegrésszel kicserélhető a két változási pont között) anélkül, hogy a prozódiai hangzásban minőségi csökkenés következne be. Ez azt eredményezi, hogy a korpuszban úgy kereshetünk szöveget, hogy meg tudjuk jósolni a dallammenetét és hangsúlyozását is.

A magyar beszéd prozódiai (szupraszegmentális) szerkezetével kapcsolatos eddigi elemzések, valamint az egyes részterületekre korlátozott modellezési formák jó kiindulási alapot szolgáltatnak a fenti vizsgálatok végzéséhez [7].

7. A prozódiai vizsgálatok anyaga és módszere

A vizsgálatokhoz olyan beszédadatbázisokat használtunk fel, amelyekben a bemeneti szövegtest mellett szerepeltettük a fonemikus átírási formát is, valamint ezzel párhuzamosan tároltuk az elhangzott beszédjelet, annak hang- és szó szintű címkéit, valamint az ezekhez tartozó időtartam, alapfrekvencia és intenzitás adatokat *(3. ábra)*. A jelen vizsgálatnál fontos szempont, hogy ugyanazon beszélő hangját vizsgáljuk és az összehasonlításokat is egyetlen hangra kell vonatkoztatni, hiszen a korpusz alapú szintézisnél is egy adott beszélő hangjából akarunk szintetizált mondatokat előállítani.



4. ábra Négy kiegészítendő kérdés egyenkénti Fo menete, és ezek egymásra helyezett képei (alul). A függőleges vékony vonalak a hanghatárokat jelölik, a vastagabb szürke vonalak pedig a szóhatárokat.

A célkitűzés megvalósítására egyszerű kijelentő és kérdő mondatok alapfrekvencia- és intenzitásszerkezetét elemeztük A mintamondatokat egy olyan beszédadatbázisból válogattuk amelyben egy férfi bemondó hangját rögzítettük [8]. A kísérlethez 16 kijelentő és 8 kérdő mintamondatot válogattunk ki az adatbázisból. A mintamondatok mindegyikén jellemeztük az alapfrekvencia változást annak töréspontjaival, valamint az intenzitások alakulását. A mondatok szövegtartalmát figyelmen kívül hagytuk a vizsgálat során, mivel mondat szinten voltunk kíváncsiak a szövegtől független Fo és I szerkezetek alakulására.

8. A prozódiai vizsgálatok eredményei

8.1. A kérdések vizsgálata

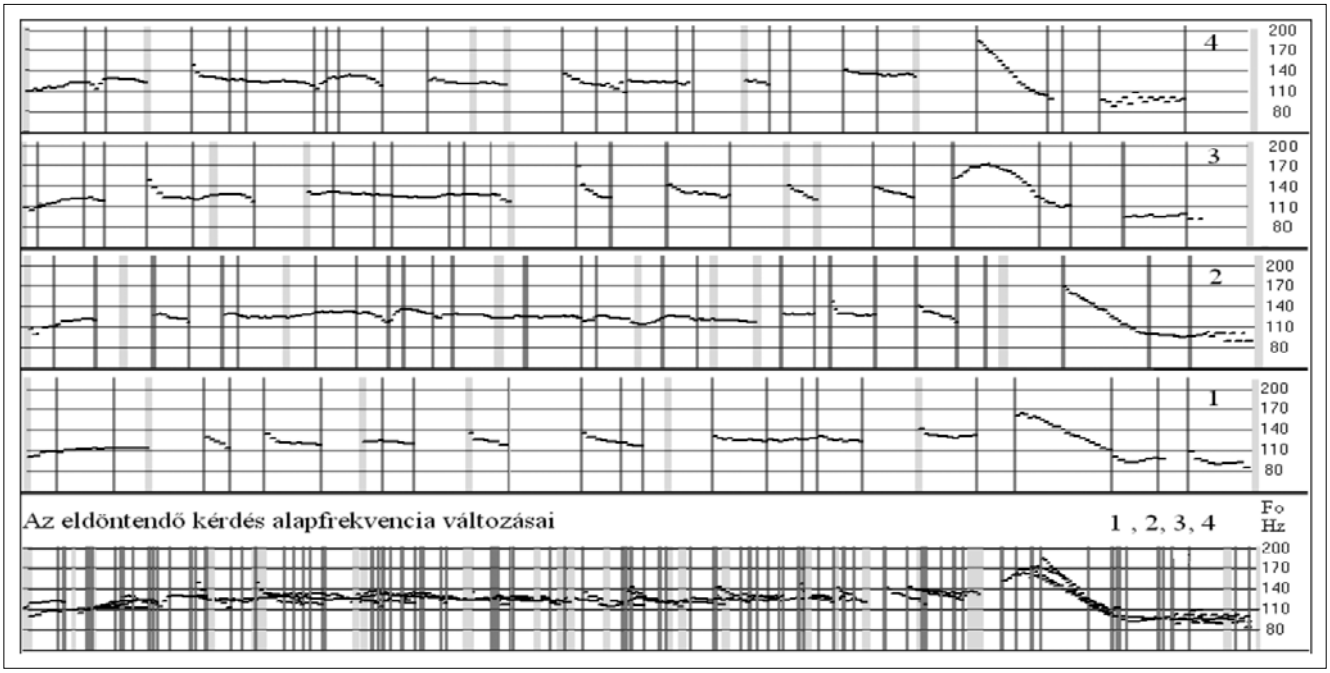
A mért kiegészítendő kérdésekben az alapfrekvencia mozgása a kérdésre jellemző jól ismert képeket mutatja, a vizsgálat tárgya itt a prozódiai szerkezet stabilitása. Az összesített vizsgálati eredmények azt mutatják, hogy a kérdésekre produkált alapfrekvencia szerkezetek egységes képet mutatnak. A kérdés magján a kérdőszón van az intonációs csúcs (meredek felfutás és meredek csökkenés), utána pedig enyhén esik az Fo.

Részletezve az előbbi leírást a vizsgált bemondóra azt mondhatjuk, hogy az Fo csúcs a kérdőszó első szótagjának magánhangzóján 200 Hz körüli értéken van. Ezután erőteljes meredek esés következik a második szótag magánhangzójának a végéig, itt 110 Hz körüli az Fo. Ettől a ponttól kezdve az alapfrekvencia folyamatosan csökken 90 Hz körüli értékre a mondat végére. A mondatok tartalmi része nem befolyásolja a dallammenetet, az ugyanazon bemondó kérdéseiből átlagolt görbe mentén kicsi az Fo szórása *(4. ábra)*. Ez azt jelenti, hogy a kiegészítendő kérdés dallammenete a vizsgált beszélőre három jellemző ponttal leírható. Az intenzitás alakulására ugyanez mondható el. Ezeket a pontokat a szöveg szintjére vetítve megkapjuk az általános sémát a kiegészítendő kérdések és a szöveg kapcsolatáról. A fenti három pont helyzetének szövegszintű azonosításához csak a szótagok helyzeti meghatározására van szükség (szó szintű információt nem használunk). A jellemző pontok a következők: az első szótag magánhangzója, a második szótagé, továbbá a mondat utolsó szótagja.

Az eldöntendő kérdésekben az Fo menet szintén stabil képet mutat, független a mondat tartalmától *(5. ábra)*. A csúcs a kérdés magján van, az utolsó előtti szótagon. Részletezve: a mondat indítására a 110 Hz körüli érték a jellemző, fokozatos emelkedés következik 140 Hz-re az utolsó előtti szótag elejéig, majd hirtelen ugrás következik 170 Hz körüli értékre amit az Fo a szótagmag elején ér el, majd a szótag végéig hirtelen csökkenés következik be a 90 Hz körüli értékre. Ez az érték marad a mondat végéig. Mint látható itt négy jellegzetes pontot lehet kijelölni a szövegtől függetlenül: a mondat kezdete, az utolsó előtti szótag eleje, ezen belül a magánhangzó eleje, majd a magánhangzó vége.

A kérdések tekintetében tehát azt az összegzett következtetést vonhatjuk le, hogy az alapfrekvencia görbe egységes, invariáns képet mutat és a szövegtől függetlenül jellemző a kérdésre.

A fentiekből látható, hogy a kérdésekben a szövegre vetített prozódiai információt a szöveg szótag szerkezetéhez lehet kötni, sem a szavak sem a szöveg tar-



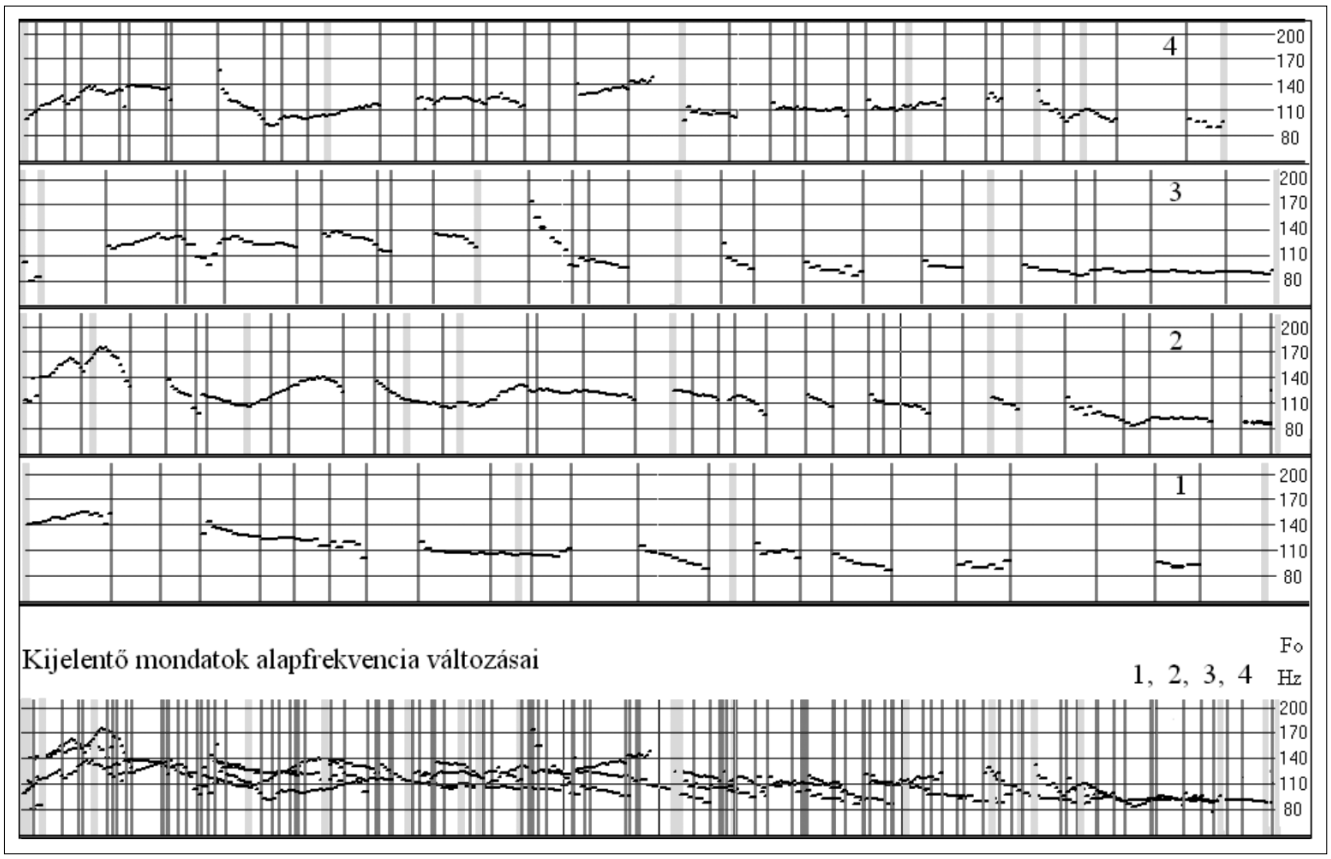
5. ábra Négy eldöntendő kérdés egyenkénti Fo menete, és ezek egymásra helyezett képei (alul)

talma nem befolyásolja a prozódiai szerkezet kialakítását. Mindezekből azt a következtetést vonhatjuk le, hogy az egyszerű kérdések előállításához nincs szükség a szöveg semmilyen elemzésére, hiszen az Fo és I töréspontok egyértelműen, szótaghoz kötötten meghatározhatók. Ez megkönnyíti a válogatást a korpuszban.

8.2. A kijelentő mondatok vizsgálata

A vizsgált mondatokra egyenként végeztük el az Fo menet elemzését és a jellemző pontok hozzárendelését a szöveghez. Az elemzések eredményei a következők. A kijelentő mondat Fo menetében változást okoz a mondathangsúly helye, a szó hangsúlyos volta, a hangsúlyos szavak helye a mondatban, valamint a határjel-

6. ábra Négy kijelentő mondat egyenkénti Fo menete, és ezek egymásra helyezett képei (alul)



zések (vessző, gondolatjel, pontosvessző, kettőspont). Úgy találtuk, hogy az Fo általában a mondat első hangsúlyos szótagján a legmagasabb értékű, de ez attól is függ, hogy a beszélő hogyan értelmezi a mondatot a hangsúlyok vonatkozásában. Amennyiben van mondathangsúly, akkor az lehet a legmagasabb értékű. A hangsúlyos szavak első szótagján Fo emelkedés található, a visszacsökkenés az esetek nagy többségében a második szótagban zajlik le. Az Fo csúcs kiemelkedésének a mértéke a hangsúlyos szótagban általában függ a szó mondatbeli helyzetétől. Minél távolabb vagyunk a mondat elejétől, annál jobban csökken az Fo kiemelkedése a környezetéből.

Az utolsó szavakban (ha ezek hangsúlyosak is) ez a kiemelkedés szinte csak pár Hz. A hangsúlyok közötti részeken az Fo enyhe esést mutat, azonban ezt az eső tendenciát megváltoztathatja a határjelzés, illetve a mondathangsúly. Ilyenkor nem eséssel kell számolni, hanem szinten tartással, esetleg enyhe Fo emelkedéssel. A kijelentő mondatok alapfrekvencia szerkezete tehát sokkal bonyolultabb képet mutat, mint a kérdéseké. A *6. ábrán* a vizsgált anyagból négy mondat Fo szerkezetét mutatjuk be.

Az összesített képből látható, hogy az Fo szórása sokkal nagyobb, mint a kérdéseknél. A mondat belsejében nem lehet jellemző Fo karakterisztikát találni. Ez a mondat belseji hangsúlyok más-más elhelyezkedéséből fakad. A mondat elejére ki lehet mondani, hogy az magasabb Fo-al rendelkezik, mint a mondat vége. Az egyedüli egységes pont, ami minden ilyen kijelentő mondatra jellemző a mondatvég alapfrekvenciájának értéke, az Fo 85-90 Hz-re csökken (a vizsgált bemondó ejtésében). Az intenzitásszerkezeti kép egységesebb, mint az alapfrekvencia. A mondat kezdetén kialakuló intenzitás jellemző a mondat nagy részére, a befejező szakaszban az intenzitás csökken.

Látható tehát, hogy a kijelentő mondatokban a prozódia és a szöveg kapcsolatának kijelölése bonyolult szövegelemzést is igényelhet, hogy a megfelelő prozódiai részeket függetlenítsük a szöveg tartalmától. Ez nem kedvező eredmény a korpusz alapú szintézis szempontjából, hiszen a legtöbb esetben kijelentő mondatokat, közléseket kell előállítani. A probléma kezeléséhez az szükséges, hogy csökkentsük a mondatok variáltságát a korpuszban. Ezért a korpusz alapú rendszereket ma még csak meghatározott témakörökre (például időjárás jelentés, jegyrendelés) fejlesztenek. Itt elérhető, hogy a korpuszban viszonylag állandó szerkezetű mondatokat tárolunk és ezekből építjük fel az új mondatot. Ennek köszönhető, hogy a nyelvi elemzésnél egyszerűbb módszerekkel is modellezni tudják a kapcsolatot a kijelentő mondat prozódiai szerkezete és a szintetizálandó mondat szövegteste között (lásd Fék és tsai. cikkében, ugyanebben a számban).

9. Összefoglalás

A korpusz alapú beszédszintézis-technológia nyelvészeti-fonetikai kérdéseit tárgyaltuk. Rámutattunk arra, hogy a beszédjel akusztikai megjelenéséhez szorosan hozzátartozik a fonetikai, nyelvi háttér is. A jó akusztikai végeredmény (emberi hangú szintetizált mondat kellemes hangszínezettel és hanglejtéssel) eléréséhez ezeket az ismereteket is fel kell használni a korpusz alapú szintetizáló rendszerek tervezésénél.

Köszönetnyilvánítás

Ezt a kutatást az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Irodalom

[1] Schweitzer A., Braunschweiler N., Klankert T., Möbius B., Sauberlich B., Restricted Unlimited Domain Synthesis. Proc. Eurospeech 2003, Geneve, pp.1321–1324.

[2] Kawai H., Toda T., Ni J., Tsuzaki., Tokuda K., Ximera: a new TTS from ATR based on corpus-based technologies. Proc. of the 5th ISCA Speech Synthesis Workshop, Pittsburgh 2004.

[3] Olaszy Gábor, Az artikuláció akusztikai vetülete – a hangsebészet elmélete és gyakorlata. Kif-LAF 2003. Szerk.: Hunyadi László, Debreceni Egyetem 2003, pp.241–254.

[4] Magdics Klára, A magyar beszédhangok akusztikai szerkezete. NytudÉrt. 49, Akadémiai Kiadó, Budapest, 1965.

[5] Olaszy Gábor, A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. NytudÉrt. 121, Bp., 1985.

[6] Tamm Anne, Olaszy Gábor, Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szerk.: Alexin Zoltán és Csendes Dóra, Szegedi Tudományegy. Informatikai Tanszékcsoport, Szeged 2005, pp.383–393.

[7] Olaszy G., The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4), 2002. pp.277–306.

[8] Olaszy Gábor, Abari Kálmán, Adatbázisok és számítógépprogramok a magyar beszéd időszerkezeti vizsgálatához. Alkalmazott Nyelvtudomány 2., 2005, pp.41–62.

Gépi tanuló algoritmus automatikus címkézésre és alkalmazása beszédszintézis céljára

KISS GÉZA, NÉMETH GÉZA

BME Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}@tmit.bme.hu

Lektorált

Kulcsszavak: automatikus címkézési módszerek, gépi tanuló algoritmus, nyelvazonosítás, LID, TTS

A cikkben egy új, szöveg címkézési problémák megoldására használható tanuló algoritmust mutatunk be. Illusztráljuk a probléma fontosságát szövegfelolvasó rendszerek, ezen keresztül távközlési alkalmazások számára. Áttekintjük a jelenleg használt módszereket a szavankénti nyelvi címkézés problémájára. Bemutatjuk az általunk javasolt módszert, és demonstráljuk eredményességét a nyelvazonosításban, három különböző tanítóhalmazra, nagyméretű tesztkorpuszokon.

1. Bevezetés

Egyre szaporodik mind hazánkban, mind a nemzetközi színtéren a beszédalapú távközlési szolgáltatások száma, amelyekben a bejövő hívásokat IVR (Interactive Voice Response) rendszer fogadja. A legmodernebb rendszerekben a felhasználó már nem csupán billentyűzéssel, hanem beszéddel is közölheti mondandóját az ASR (Automatic Speech Recognition) rendszeren keresztül, és válaszul jó minőségű beszédet kap. Egyszerű esetekben a válasz előre felvett (prompt) illetve néhány elemből összeállított (kötött szótáras beszédszintézis) hangüzeneteket tartalmaz.

Amennyiben a kimondandó üzenet tartalma megjósolhatatlan, vagy nagyon nagy variabilitást mutat, akkor a beszédet szövegfelolvasó rendszerrel (TTS, Text-to-Speech) állítjuk elő. Néhány példa az utóbbira a hazánkban működő szolgáltatások közül: a T-Mobile telefonos e-mail felolvasó rendszere [1], számszerinti tudakozó rendszere [2], vagy a T-Com hangos SMS szolgáltatása.

A TTS rendszerektől elvárjuk, hogy pusztán az írott alakból jó minőségű, érthető és helyesen intonált beszédet hozzanak létre. Azonban elmondható, hogy a használt írásrendszerek (akár a magyar, akár más nyelvé) a beszéd információtartalmának csak egy tört részét tartalmazzák, a prozódiára csupán kis mértékben, néhány írásjeggyel utalnak. Az ember, ha olvas, a világról való ismerete valamint a szövegkörnyezet alapján egészíti ki elméjében a szöveget a hiányzó információkkal, ami segíti abban, hogy (szükség esetén) azt megfelelő kifejező erővel fel is olvassa. Pusztán a helyes kiejtés, hangsúlyozás megállapításához is szükség van olyan információk helyes felismerésére, mint a szöveg nyelve, az esetlegesen a szövegbe beékelt idegen szavak nyelve, az egyes szavak szerepe a mondatban (szófaj, nyelvtani szerkezet).

Az *1. ábrán* láthatunk néhány példát, ahol ez nem triviális. Egyező szóalakok esetén a helyes eredmény eldöntése nehéz, illetve a mondat nyelvétől eltérő nyelvű szó beékelődése esetén is.

A cikkben áttekintjük a szöveg alapú nyelvazonosításra használt módszereket, majd leírunk egy felcímkézett szövegből történő tanulásra szolgáló gépi tanuló algoritmust, amely használható különböző címkézési feladatok automatikus elvégzésére. A módszert nyelvazonosításra történő címkézés példáján mutatjuk be, utalva a szófaji címkézés lehetőségére is, amelyek például a távközlésben is egyre gyakrabban használt TTS rendszerek számára fontos információk; ezek mellett a módszer feltehetően számos más területen is használható.

2. Módszerek szöveg alapú nyelvazonosításra

A nyelvazonosítás megnevezéssel (Language Identification, LID) egyaránt illetik az ASR alkalmazása előtt a beszéd nyelvének megállapítására vonatkozó módszereket, és a szöveg nyelvének megállapítására vonatkozókat. A nyelvazonosítás tekinthető a címkézési, illetve osztályozási problémák egy speciális esetének, így a tanulságok más területen is alkalmazhatók - például szófaji címkézés esetére is.

Bár első ránézésre a legtöbb, a téma iránt kicsit is érdeklődő embernek vannak ötletei írott szöveg nyelvének automatikus megállapítására, valójában több olyan kérdés nehezíti a megoldást, ami miatt a feladat közel sem triviális. Míg viszonylag egyszerű módszerekkel nagy valószínűséggel a helyes nyelvet rendel-

1. ábra Példák nem triviális nyelvi illetve szófaji címkézési feladatokra

„a test” – magyar vagy angol kifejezés:	A lélek és a test.	This is a test.
idegen kiejtésű rész magyar szövegben:	A „Sok hűhó semmiért” Shakespeare műve.	
„egy” – számnév vagy határozatlan névelő:	Egy vagy két alma.	Egy alma esett le a fáról.

hetjük egy hosszabb szöveghez, a szöveg szavai nyelvének helyes megállapítása kevert nyelvű, több nyelvben is előforduló szóalakokat tartalmazó szövegen jóval nehezebb feladat, kiegészítve azzal, hogy a valóban hatékony megoldásnak nem csak pontosnak, hanem gyorsnak és viszonylag kis tárigényűnek kell lennie. Valamennyi követelménye egyidejű teljesítése már összetett, nehezen megoldható feladat.

2.1. Morfológiai elemzés

A módszerek egy csoportja a szószinten, sőt morféma-szinten való helyes azonosításra törekedve részletes morfológiai elemzést alkalmaz, például a DCG-k (Definite Clause Grammar) használatával [3], esetleg közvetve egy helyesírás-ellenőrző használatával [4]. Egy más, köztes megoldásban nem történik valódi morfológiai elemzés, hanem szótárak (szó és szóelem-listák) elemeire való illeszkedés alapján következtetnek a szavak nyelvére, kiegészítve ezt statisztikai módszerekkel [5]. Magyar nyelven elérhető morfológiai elemző a Humor [6], valamint a Hunmorph [7]. Azonban ha nem csak a magyar/nem magyar döntést kell meghoznunk, hanem több lehetséges nyelv közötti döntés is szükséges, akkor mindegyikhez szükséges morfológiai elemző, ami nehezen kivitelezhető feladatot jelent, valamint egyes alkalmazásokban problémát okozhat a szükséges nagy számítási kapacitás.

2.2. Szóalapú módszer

A szóalapú módszerek [8] azon a megfigyelésen alapulnak, hogy minden nyelvben van a szavaknak egy olyan, viszonylag kis halmaza, amelyet nagyon sokat használnak. Ezért egy nyelvhez tartozó ilyen szavak jelenléte nagy biztonsággal jelzi, hogy a szöveg az adott nyelven íródott. A leggyakoribb 1000 szó az összes előforduló szó 50-70%-át is kiteheti [9].

2. ábra
Példák öt európai nyelv leggyakoribb szavai között többen is előforduló szóalakokra

szóalak	angol	német	spanyol	lengyel	magyar
de			X		X
a	X		X		X
na				X	X
to	X			X	
in	X	X			
do	X			X	
el			X		X
is	X				X
es		X	X		
mit		X			X
was	X	X			
ha			X		X
so	X	X			
on	X			X	
mi			X	X	X
most	X				X
ja		X		X	
be	X				X
ma				X	X

A módszer hátránya, hogy minél rövidebb a szöveg, annál valószínűbb, hogy a halmaz egy szava sem jelenik meg benne. Emellett a szólista összeállításához is számottevő erőfeszítés kellhet, hiszen vannak olyan szavak, amelyeknek az írott formája több nyelven is ugyanaz, mint ahogy azt a 2. ábrán található példákból látjuk.

2.3. Vektortér módszerek

A vektortér módszerek alapgondolata, hogy a vizsgálandó dokumentumhoz és a lehetséges besorolási kategóriákhoz is egy-egy jellemzővektort rendel, amelyek bizonyos számszerűsíthető tulajdonságok leképzései, a kategóriák esetében a jellemző fontosságával súlyozva. A dokumentumnak az egy kategóriába való illeszkedését a jellemzővektorok skalár-szorzatával jellemzi, ahol a 0 érték ortogonalitást, az 1 érték megfelelést jelent. A [10]-ben leírt módszerben a jellemzők N-gramok N=2..5 értékekkel, valamint rövid, illetve korlátlan méretű szavak, melyekkel szövegek nyelvének megállapítását végzik.

2.4. Neurális hálók

A [11] megközelítésében egy többrétegű perceptront (Multi Layer Perceptron, MLP) tanítanak be, amelynek bemenetére a szöveg egy pozíciójára illesztett ablakba eső karaktereket helyezik, kimenetén pedig nyelv-valószínűség értéket ad. Ezeket a szó összes betűjére kombinálva szavankénti nyelvi döntést hoznak.

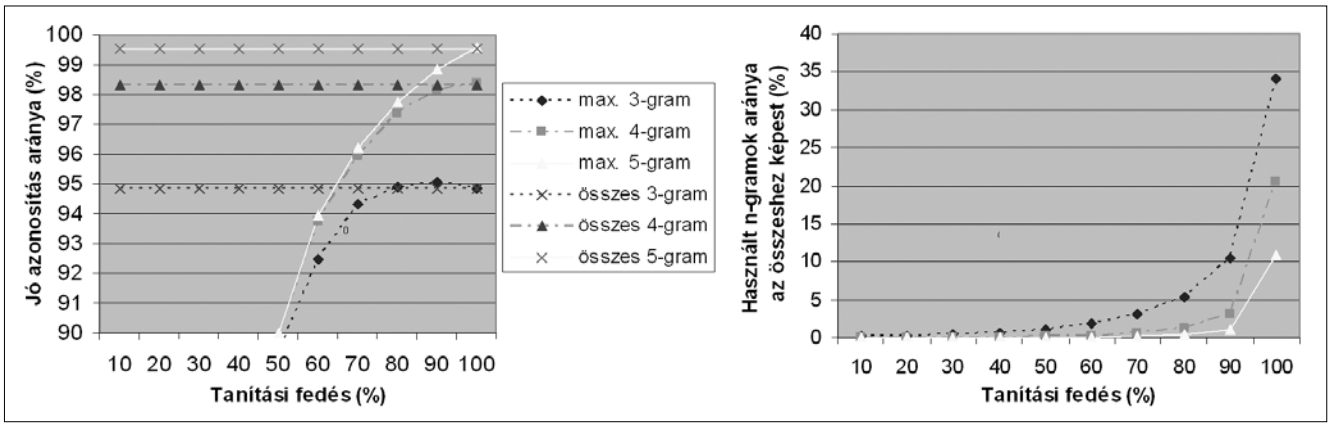
2.5. Döntési fa alapú módszerek

A 3. pontban javasolt módszerünktől eltérő döntési fa tanítást használ [12]: karakterenként külön döntési fát tanítanak, ahol az ágak a szomszédos karakterek azonosságára kérdeznek rá, a leveleken pedig a legvalószínűbb nyelv címkéje található. Szavanként hoznak döntést a szó karaktereire kapott nyelv-jelöltek közül a legtöbbször előfordulóra döntve.

2.6. N-gram alapú módszerek

Az N-gram alapú módszerek szórészeket használnak az azonosítás alapegységéül. Ezek két, három vagy több karakter sorozatából állhatnak; egy módszeren belül esetlegesen több különböző hosszt is használhatunk egyidejűleg. Az N-gram gyakoriságok statisztikáját a különböző nyelvekhez tartozó tanító szövegkorpuszból készíthetjük.

Az N-gramok jól kezelnek több problémát, amelyekben a szóalapú módszerek nem adnak megoldást. Ezek egyike az elektronikus szövegekben gyakori betűhibák (elgépelések, karakterfelismerési hibák), mivel ezeknek olyan nagy változatossága van, amit nem vagy nagyon nehezen lehet megoldani a szóváltozatok tárolásával, viszont az n-gram statisztikát



3. ábra
A rögzített és változó hosszúságú előzményen alapuló módszerek teljesítményének és méretének összehasonlítása a tanító halmazon ML becslés használatával

nem rontják el nagy mértékben. Egy másik szempont, hogy a „data sparseness" probléma (amely szerint gyakorlatilag soha nincs annyi adatunk, amely minden szóbajöhető elemről adna eloszlási információt; mindig találkozunk majd a tanítóhalmazban elő nem fordulókkal) jóval kisebb mértékben jelentkezik ebben az esetben, mintha szavakat vizsgálnánk, mivel egy szóban a szóhossz négyzetével arányos számú N-gram található. Egy jellegzetes példája ennek a megközletésnek Canvar és Trenkle módszere [13].

2.7. Markov-modell

Tudjuk, hogy egy *l* karakter hosszú szó előfordulási valószínűségét megkaphatjuk a láncszabály szerint:

$$\mathrm{P}(\text{szó} \mid \text{nyelv}) = \prod_{i=1}^{l+1} \mathrm{P}(c_i \mid c_0 \ldots c_{i-1}, \text{nyelv}) \qquad (1)$$

$c_1,\ldots,c_l$ ahol a szó karakterei, $c_i, i \leq 0$, illetve c_{l+1} speciális, szót kezdő, illetve bezáró jelek.

Ezt a valószínűséget közelíthetjük a Markov-modell segítségével, azaz feltételezve, hogy ez egy véletlen folyamat, amelyben a következő karakter valószínűségi eloszlása csak a jelenlegi állapoton múlik. Hagyományosan az állapoton az előző *n-1* karaktert értik a nyelvazonosítás esetén:

$$\mathrm{P}(\text{szó} \mid \text{nyelv}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \mid c_{i-n+1} \ldots c_{i-1}, \text{nyelv}) \qquad (2)$$

A karakterek egy adott környezethez tartozó feltételes valószínűségét az ML (Maximum Likelihood) becslés alapján közelíthetjük:

$$\mathrm{P}(c_i \mid c_{i-n+1} \ldots c_{i-1}) \approx \frac{\#\, c_{i-n+1} \ldots c_{i-1} c_i}{\#\, c_{i-n+1} \ldots c_{i-1}} \qquad (3)$$

Mivel gyakorlatilag bármekkora méretű tanítóhalmaznál kell számítanunk arra, hogy előfordulhatnak korábban nem látott N-gramok, ezért elkerülhetetlen valamilyen simító (smoothing) módszer használata, amelynek megválasztása jelentősen meghatározza a becslés minőségét.

A szó nyelven belüli valószínűségét ezután használhatjuk a nyelv valószínűségének becslésére:

$$\mathrm{P}(\text{nyelv} \mid \text{szó}) = \frac{\mathrm{P}(\text{szó} \mid \text{nyelv}) \cdot \mathrm{P}(\text{nyelv})}{\mathrm{P}(\text{szó})} \qquad (4)$$

A legvalószínűbb nyelv megállapításához a nevezőt (amely a vizsgált szövegre, és nem a nyelvre jellemző érték) figyelmen kívül hagyhatjuk. A nyelv valószínűségét vehetjük fix értéknek, vagy a kontextus alapján dinamikusan változónak.

$$\text{nyelv} = \arg\max_{\text{nyelv}} \mathrm{P}(\text{szó} \mid \text{nyelv}) \cdot \mathrm{P}(\text{nyelv}) \qquad (5)$$

Elméletileg a Markov-modellel való becslés megfelelő simító módszer választásával tetszőlegesen pontos becslést adhat, ha *n* elég nagy. (Ha *n* a maximális szóhossz, akkor a láncszabályt, ezen keresztül a pontos valószínűséget kapjuk.) A gyakorlatban azonban rendszerint $n = 2$ (bigrammok) vagy $n = 3$ (trigrammok) értéket használnak, két okból: a data sparseness probléma miatt, és mert nagy *n*-hez számottevő tárolási kapacitásra volna szükség.

Azonban ezek a hosszak nem teszik lehetővé a pontos osztályozást több nyelvre való döntés esetén, ahogy ezt a *3. ábrán* láthatjuk ML becslés használatával az *1. táblázat* első sorában leírt tanítóhalmazra; ez az elvi korlátot jelenti. Mint láthatjuk, az adott tanítóhalmazon való tanítás és ugyanazon való tesztelés esetén a helyes azonosítási arány még 5-grammok esetén sem éri el a 100%-ot, de ez már meglehetősen nagy adatbázist jelentene, valamint az ismeretlen szövegen feltétlenül használandó simítást még nem tartalmazza. A javasolt módszerrel létrehozott adatbázis mérete az előző 10-35%-a hasonló felismerési aránynál, valamint 100%-ra való tanításnál csekély mértékben jobb eredményt is ad.

2.8. Tanulságok a nyelvazonosításról

Ebben a szakaszban a szövegből történő nyelvazonosítás néhány módszerét tekintettük át, melyeknek két nagy csoportja a részletes elemzés alapján való döntés és a statisztikai módszerek.

Összefoglalásként elmondható, hogy a jelenleg használt, tisztán statisztikai alapú megközelítések általában nem adnak eléggé pontos nyelvazonosítást rö-

vid szövegeken, így a szavak szintjén való azonosításhoz nem elég megbízhatóak. Emellett azok, amelyek a dokumentumot előzetesen tanító szövegtörzsből nyelvenként készített „nyelvi profilokhoz” hasonlítják a vizsgálandó részt, gyakran számottevő számítási kapacitást igényelnek az azonosítási fázisban is. A részletes morfológiai elemzés végzése viszont nehezen kivitelezhető, főként nagyszámú nyelvre, valamint problémát okozhat egyes alkalmazásokban a szükséges nagy számítási kapacitás.

3. A javasolt módszer

Az alábbi módszert elsősorban a szövegből történő nyelvazonosítás feladatára dolgoztuk ki, ezért ebből a szempontból tárgyaljuk. Azonban a „nyelv” kifejezés helyett mindenütt „osztályt”, a „szó” helyett „szövegegységet” behelyettesítve, egy általános címkéző rendszer leírásaként is olvashatjuk.

3.1. A módszer alapelve

A célunk olyan megoldás kidolgozása volt, amely lehetővé teszi nagyon rövid szövegek helyes azonosítását is, akár a szavak szintjéig, és amely kézben tartható abban az értelemben, hogy be lehet tanítani tetszőleges bemenet helyes azonosítására, de egyben általánosító képességgel is rendelkezik, azaz nem látott szavak nyelvét is képes helyesen felismerni a tanítóhalmaz szavaihoz való hasonlóság alapján. Emellett célunk volt a működéshez szükséges adatbázis méretének korlátok között tartása is.

Ennek a célnak megfelel, ha a P (szó | nyelv) valószínűséget egy előzetesen rögzített kritériumnak megfelelő pontossággal becsüljük meg – például elég pontosan ahhoz, hogy arra a nyelvre legyen a legnagyobb a becsült valószínűség, amelyre legnagyobb a tanítóhalmazból számolt valószínűség. Módszerünk másik összetevője, hogy a szavak kontextusa alapján számítjuk az adott szóra a nyelv-valószínűséget. Ezután az (5) egyenletnek megfelelő nyelvre döntünk minden szó esetén.

Ez a megközelítés elvileg lehetővé teszi, hogy szószinten helyes nyelvazonosítást kapjunk, még homomorf (esetünkben egyidejűleg több nyelvhez tartozó) szavak esetén is a szövegkörnyezetnek megfelelően, valamint hogy egynyelvű szövegbe beszúrt idegen nyelvű szó a valódi nyelvének megfelelő azonosítást kapja, szemben a környezet alapján determinisztikusan döntő naiv megközelítéssel. Ha szövegrészenként (pl. mondatonként) egy nyelvre való döntés szükséges, a szavakra meghatározott nyelvi címkék alapján dönthetünk, például a többségi szavazás elvének megfelelően.

Megfelelő valószínűség-becslési módszerrel az ismert szavak írásmódja alapján képesek lehetünk korábban nem látott szavakra is becsülni ezt a valószínűséget. Ez a megközelítés megőrzi a szó-alapú módszerek előnyét, a kézbentarthatóságot, kiterjesztve azt általánosító képességgel, és szóalapon is helyes működést tesz lehetővé. A megközelítés sikerességéhez a kulcs a feltételes valószínűség és a nyelvi valószínűség megfelelő pontosságú becslése.

3.2. Feltételes valószínűségek becslése

A P(szó | nyelv) valószínűség becslésére általunk kidolgozott módszer változó méretű N-gramok használatán alapszik. Míg a szokványos Markov-modellt használó megoldásban rögzített hosszúságú előzményt használunk egy karakternek az előzőek után való következésének valószínűségi becslésére, a javasolt módszerben többféle hosszúságú előzményt használunk.

$$\mathrm{P}(\text{szó}\,|\,\text{nyelv}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \,|\, c_{i-n_i+1}\ldots c_{i-1}, \text{nyelv}), n_i \geq 0 \quad (6)$$

Az n_i hosszt minden környezetre egy tanítási folyamat során határozzuk meg. A tanítás 0 hosszúságú karakter-környezettel indul minden karakterre (ez a karakter előfordulásának valószínűsége), majd ezt a hosszt bizonyos környezetekben növeli a megcélzott valószínűség-becslési kritérium elérésére, amely lehet például a leggyakoribb szavak bizonyos százalékának helyes felismerése. A növelési folyamat korlát nélküli folytatása a láncszabályt adja, ezzel pedig nyelvenkénti szó-valószínűséget (megfelelő simítómódszer alkalmazása esetén), ezért a tanító folyamat tetszőleges tanító halmaz esetén jobb szó-valószínűség becsléshez, így a tanítóhalmazra a korrekt azonosításhoz konvergál. Hosszabb N-gramokat tartalmazó, nagyobb méretű felismerő adatbázis használatával pontosabb azonosítási eredmény érhető el. Ezáltal a módszer skálázható, mivel az adatbázis méret-kívánt felismerési arány-páros egyike szabadon megválasztható.

Az N-gram környezetekhez tartozó feltételes valószínűségeket fában tárolva úgy is felfoghatjuk a módszert, hogy egy fajta döntési fa tanítását jelenti a szóvalószínűségek becslése céljából. A fa bővítésének irányát a bővítésnek a becslési kritérium szempontjából meghatározott „hasznossága” szerint határozzuk meg. Több ilyen hasznosság-függvénnyel dolgoztunk, melyeket a tanító halmazon való helyes felismerési arány és az attól független teszt-halmazon való eredmény, azaz az általánosító képesség mellett az alapján is vizsgáltunk, hogy mennyire tömör adatbázist eredményez. A tömörséget nem pusztán a mérettel jellemeztük – hiszen nem közömbös, hogy milyen felismerési arányt ad a tömörebb adatbázis – hanem a felismerési arány/méret hányadossal, amelyet a LID-adatbázis teljesítményének nevezünk. A legjobb hasznosság-függvényeket és az őket jellemző grafikonokat a *4. ábrán* láthatjuk. A legjobb általánosító képességű megoldáshoz és a legtömörebb adatbázishoz eltérő hasznosság-függvény szükséges.

Újítás még, hogy a nyelvek független szemlélése helyett a nyelvenkénti valószínűségek eltérésének elégségesen pontos becslésére törekszünk, amelytől kisebb adatbázis méretet várunk, hiszen így a tanítás során a két nyelvet megkülönböztető jellemzőkre való

„koncentrálásra” késztetjük az algoritmust. Ezáltal az algoritmus nem csak skálázható, hanem automatikusan skálázódik is a probléma nehézségének megfelelően. Például, ha két nyelvet a karakterkészlete is megkülönböztet, akkor a 100%-osan helyes azonosítás megcélzásakor is megáll a tanítás az unigrammok (1 karakteres N-gramok) használatánál.

3.3. Nyelvi valószínűségek használata

3.3.1 A nyelvi valószínűség fogalma

A „nyelvi valószínűség” kifejezésen az egyes nyelvek bizonyos környezetben való előfordulásának valószínűségét értjük; ezt használjuk (5)-ben a P(nyelv) helyén. E valószínűség becsült értékének kiszámításában megoldásunkban részt vehet a környező szavak nyelve és a közöttük lévő központozás. Azért választottuk ezeket a lehetséges vizsgálandó jellemzők közé, mert ezek szerepet játszhatnak az emberi értelmezés kialakításában is, de a lehetséges kérdéseknek más, szűkebb, illetve tágabb halmaza is elképzelhető.

A nyelvi valószínűség modellezését is tekinthetjük egy döntési-fa tanításnak, amelyben a 3.2 pontban adotthoz hasonlóan feltételes valószínűségek tárolására használjuk a fát, ám attól (és az általában használt módszerektől [15,16]) eltérően, nem csupán a szót megelőző, hanem az azt követő szavak azonosságát is felhasználhatjuk.

Ebben a megközelítésben nehézséget jelent egyrészt az, hogy a szó környezetében lévő szavak nyelvét is ismertnek tételezzük fel a szó nyelvének megállapításához, ami a valóságban nem teljesül, másrészt az, hogy a sok lehetséges kérdésfajta miatt, amelyek a fa ágain való továbbhaladást vezetik, nagyon nagy lehet a lehetséges döntési-alternatívák száma. Ezeknek a megoldását a következő két pontban tárgyaljuk.

3.3.2 A legvalószínűbb címke-sor megkeresése

Az első nehézség egy matematikailag megfogható problémát takar, bár a hatékony megvalósítás nem triviális. A feladat az, hogy megtaláljuk a címke-sort, amelyre az N szóból álló mondatra a valószínűség maximális lesz.

$$\{\text{nyelv}_{l_i} \mid i \in [1..N]\} = \arg\max \prod_{i=1}^{N} \text{P}(\text{nyelv}_{l_i} \mid \text{szó}_i) =$$

$$= \arg\max \prod_{i=1}^{N} \frac{\text{P}(\text{szó}_i \mid \text{nyelv}_{l_i}) \cdot \text{P}(\text{nyelv}_{l_i})}{\text{P}(\text{szó}_i)} \quad (7)$$

A $P(szó_i)$ tényezőt itt is figyelmen kívül hagyhatjuk, hiszen az nem függ a nyelvi címkétől, így az eredményt nem befolyásolja.

A legvalószínűbb címkesor kimerítő kereséssel történő megkeresése L számú lehetséges címke esetén L^N számítási lépést jelentene, ami a mondat hosszával exponenciálisan növekvő keresési teret határoz meg; ez kevés gyakorlati alkalmazásban megengedhető. Ezért az optimálist valamilyen módon közelítő módszer alkalmazása szükséges.

A jelenlegi rendszerben az alkalmazott közelítés a következő: először uniform nyelv-valószínűséggel kiszámítjuk a címkesor egy közelítését, majd az így kapott környezetet figyelembe véve újraszámítjuk a címkéket az egész sorra, balról jobbra haladva. Ha egy címke módosul a korábbihoz képest, akkor a tőle balra eső és általa (mint környezet által) esetleg befolyásolt címkéket újraszámítjuk, de csak akkor módosítjuk (részben az iteráció elkerülése végett) ha a szó nyelvére számolt valószínűség a korábbi nyelvnél nagyobb. A módszer tovább javítható, például szimulált lehűtés alkalmazásával (csökkenő mértékben engedve közvetlenül nem valószínűség növekedést eredményező címke-módosításokat is).

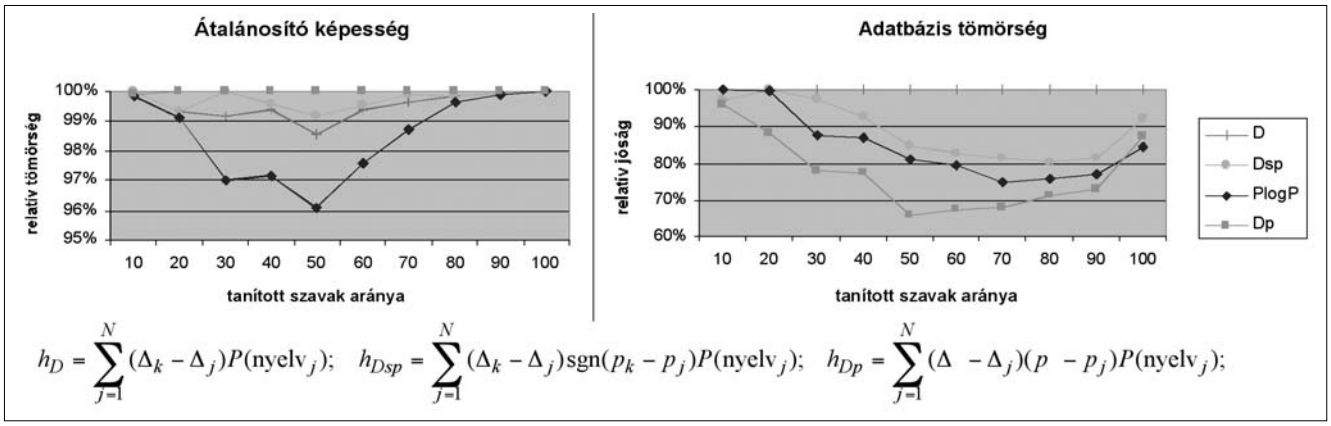
3.3.3 Szabály-sablonok használata

A második nehézség kiküszöbölésére mintarendszerünkben úgynevezett szabály-sablonokat alkalmazunk, amely teret ad a nyelvészeti tudás használatának, illetve heurisztikának, és ezzel számottevően csökkentheti a megvizsgálandó alternatívák számát.

A szabálysablonok a következő formát vehetik fel (BNF leírásban):

<sablon> ::= { <címke-leírás> [<szeparátor-leírás>] }
<címke-leírás> ::= L{[<azonosító szám>][?|*] | "<címke-név>"}
<szeparátor-leírás> ::= S{[<azonosító szám>][*] | "<szeparátor>"}

4. ábra Hasznosság-függvények jellemzése
(a legjobb eredményhez viszonyított értékek a leggyakoribb szavak különböző százalékaira);
L a nyelvek száma, k a szó valódi nyelve, p a felveendő n-gram előfordulásának valószínűsége,
Δ az n-gram felvétele miatt a számolt valószínűség-értékben bekövetkező változás.



$$h_D = \sum_{j=1}^{N} (\Delta_k - \Delta_j) P(\text{nyelv}_j); \quad h_{Dsp} = \sum_{j=1}^{N} (\Delta_k - \Delta_j) \text{sgn}(p_k - p_j) P(\text{nyelv}_j); \quad h_{Dp} = \sum_{j=1}^{N} (\Delta - \Delta_j)(p - p_j) P(\text{nyelv}_j);$$

A címke ebben az esetben a nyelv Az egyforma azonosító számmal rendelkező címkéknek illetve szeparátoroknak meg kell egyezniük. A csillaggal („*”) jelölt leírások olyan elemeket jelölnek, amelyeknek a különböző értékei nem hoznak létre egymástól független szabályokat, csupán az egymás közötti (közös azonosítóval rendelkezők közötti) egyezésnek kell teljesülni, míg a csillaggal nem jelölt leírások a címke vagy szeparátor minden előforduló értékére külön szabályt hoznak létre.

A kérdőjellel („?”) jelölt címke-leírás az, amelyre az egyes címkék előfordulási valószínűségeit megfigyeljük a tanító halmazban. A használandó szabályok előállításához a szabálysablonokat ennél a megfigyelési pontnál fogva illesztjük a tanítóhalmaz minden szavára, és ahol a sablon illeszthető, a konkrét illeszkedő értékekkel kitöltött szabályt hozunk létre, amely a folyamat végén tartalmazza az egyes címkék előfordulási valószínűségét (az illeszkedések számából, és azon belül az egyes címke-fajták előfordulásának számából számítva).

Ha egy pozícióban több szabály is alkalmazható, akkor eltérő számú feltétel-résszel rendelkezők esetén a több feltételt tartalmazót alkalmazzuk (a döntési fa elvét követve). A szabálysablonokra és az azokból előállított szabályokra később láthatunk példákat az *5. ábrában*, a nyelvazonosítás témakörére.

4. Alkalmazás szöveg alapú nyelvazonosításra

4.1. Tanítóhalmaz gyűjtése

Elmondható, hogy jelenleg nem állnak rendelkezésre nagyobb méretű, szószinten helyes nyelvi címkékkel ellátott szöveganyagok, valamint valóban egynyelvű nagy méretű szövegtörzsek összeállítása is nehéz (már csak a minden nyelvben használatban lévő idegen eredetű nevek és kifejezések miatt is, melyek olyan, egy nyelvre specializált korpuszokba is bekerülhetnek, mint például a Project Gutenberg). Ezért a tanításhoz használható szöveghalmazok összeállítása is nehézségekbe ütközik, annak ellenére, hogy az Internetről nagyon nagy mennyiségű szöveg tölthető le gyakorlatilag tetszőleges nyelvre – melyek persze az említett kevert jelleget mutatják.

A probléma egyik áthidalása lehet, hogy a szövegeket egynyelvűnek tekintve betanítjuk a nyelvazonosítót több nyelvre, majd ezzel címkézzük a szövegeket. A tanító korpuszok névleges nyelvétől egyértelműen eltérő nyelvűnek megállapított szövegeket, illetve mondatokat kihagyva, a tanítás ismételhető, így már tisztább, az egynyelvűt jobban közelítő szöveggel taníthatjuk újra a rendszert. A folyamat ismételhető, amíg történik finomodás.

Viszonylag rövid méretű szövegeket tartalmazó korpuszon (ilyen lehet például egy újság cikkeinek archívuma) jelentősen torzíthatja a szó- illetve n-gram statisztikákat a szövegekben általában jelenlévő, nagyrészt ismétlődő fej- és láblécek miatt, hiszen ezek néhány (esetleg egyébként ritka) szó, kifejezés előfordulásainak számát megsokszorozzák. De az elvileg helyes működés érdekében hosszabb szövegek esetében is érdemes ezeket a szövegrészeket eltávolítani. Ez ahhoz is hozzájárul, hogy az idegen nyelvű összetevőktől való megtisztítás hatékony legyen, hiszen idegen szövegek esetében is általában a korpusz nyelvének megfelelő nyelvű fej- és lábléc található a fájlokban.

Külön odafigyelést igényel, hogy a gyűjtött szövegek karakterei többféle kódkészlettel lehetnek kódolva ugyanazon nyelv esetén is. A tanító szövegtörzs esetében szükséges a kódkészlet ismerete, hogy ennek megfelelően kezeljük. Ha várhatóan a felismerendő bemenet kódkészletét is ismerjük, akkor csak a közös kódkészletbe (célszerűen unicode) való konvertálásról kell gondoskodni. Ha nem ismert, akkor különböző kódolású szövegekkel taníthatjuk a rendszert, vagy a másmás kódolású tanítóhalmazokat eltérő nyelvűnek véve taníthatjuk a nyelvazonosítót, így a nyelv azonosításával egyidőben megtörténik a kódkészlet azonosítása is.

További probléma az, hogy egyes alkalmazási területeken (SMS-ek, e-mailek nyelvének meghatározása) az ékezetek hiányozhatnak a szövegekről, ami megzavarhatja a nyelv-azonosítást, ha nem szentelünk neki figyelmet. Lehetséges megközelítések a mindkét jellegű szöveget tartalmazó tanító halmaz használata, illetve a tanító szöveghalmaz és a felismerendő szövegek szűkebb (ékezet nélküli) karakterkészletre való konvertálása, a nyelv azonosításakor pedig a szó eredeti karakterkészlete alapján a számított nyelv-valószínűség módosítása [11].

4.2. Az alkalmazott teszthalmaz

Több tesztet végezünk eltérő méretű tanító és felismerendő szövegkorpuszon. Először három nyelvre (angol, német, magyar) végeztünk betanítást nagyméretű korpuszon (British National Corpus, Project Gutenberg DE, Magyar Elektronikus Könyvtár), azoknak a hozzávegyült idegen nyelvű részektől való megtisztítása nélkül, a leggyakoribb szavak 90%-ának helyes fe-

1. táblázat Eredmények különböző tanító halmazok esetén, egy azoktól független 3 nyelvű teszt szöveggel, szó és mondat szintű azonosításra

Tanító (szavak)	Nyelvek	Adatbázis	Tanító, szó	Teszt, szó	Teszt, mondat
2-9 millió szó	3	54 kbájt	99,6%	94,2%	98,5%–99,5%
600-700 szó	3	7,4 kbájt	95,5%–97,8%	79,6%–87,4%	91,7%–97,2%
500-1700 szó	77	5,4 Mbájt	70,0%–99,8%	30,1%–59,6%	71%–84,0%

szabály-sablonok

L1* S* L? S* L1*

L1* S* L? S* L2*

a tanító korpuszból kapott szabályok

"": az összes alkalmazási lehetőség száma; "L": az egyéb címkék száma (sem L1, sem L2)

L1* S* L? S* L1*; { "": 13300305, "L1":13230575, "L":69730 }

L1* S* L? S* L2*; { "": 20188476, "L1":16625250, "L2":16625298, "L":168503 }

5. ábra A használt szabálysablonok és a kapott szabályok

lismerésére. A tesztet az előzőtől független szöveghalmazon végeztük (Project Gutenberg, online magyar újságok). Az [13]-ban bemutatott módszer egy web-en megtalálható implementációjához (http://odur.let.rug.nl/~vannoord/TextCat/Demo) használt, 77 nyelvhez tartozó kis méretű (5 kilobájt) szövegre is elvégeztük a betanítást.

4.3. Eredmények nyelvi valószínűség használata nélkül

A helyes azonosítás százalékos eredményeit az *1. táblázat* tartalmazza. Az osztályozott szövegek áttekintése azt mutatta, hogy az első esetben a más nyelvűnek osztályozott szövegek gyakran valóban nem a csoportjuknak megfelelő nyelvhez tartoztak, vagy kevert nyelvűek voltak, valamint hogy a valóban pontos szó-alapú működéshez szükség van egyes (formátumukat tekintve) nyelv-függetlennek tekinthető kifejezések külön beazonosítására, melyekre példák a római számok, Internet és e-mail címek, dátumok, nemzetközi szavak (pl. „tel.", „fax."), rövidítések, mértékegységeket tartalmazó kifejezések (pl. „2 cal").

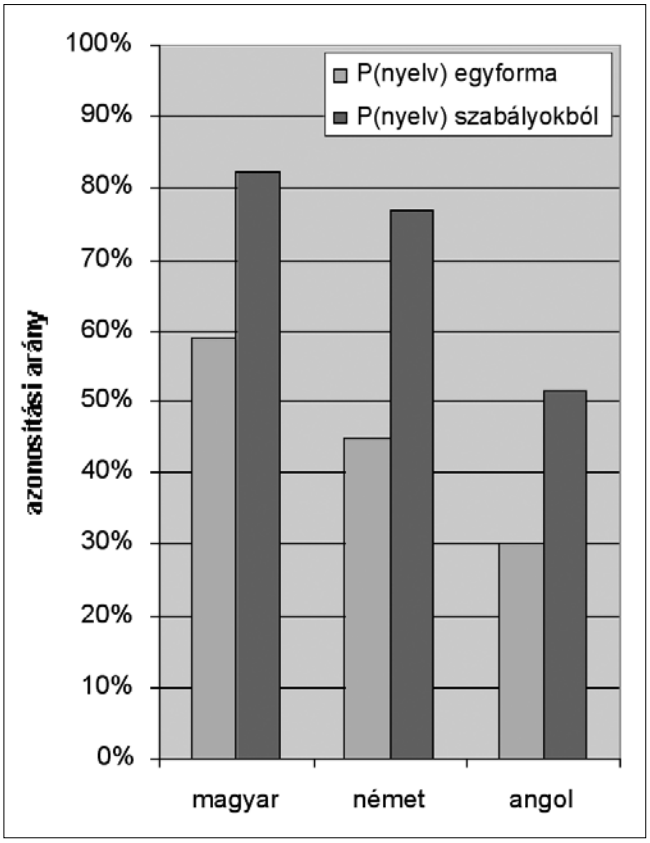
4.4. Eredmények nyelvi valószínűség használatával

Vizsgáltuk a szavak környezete alapján számított nyelv-valószínűség figyelembevételének hatását is. Ehhez az azonosító által nyelvi címkékkel ellátott szövegre alkalmaztuk a *5. ábrán* látható szabály-sablonokat, és az ugyanitt látható szabályokat. A P(szó | nyelv) valószínűségeket az 1. táblázat 3. sorára vonatkozó szöveges adatbázisból származtattuk, mivel a kis tanítóhalmaz és az ebből fakadó viszonylag gyengébb azonosítási arány nagyobb kihívást jelent a módszernek. A P(nyelv) értéket a kapott szabályokkal számoltuk, a 3.3.2 pontban leírt közelítő módszerrel.

Ezzel például a német korpuszon a korábbi 45%-ról 65%-ra növekedett a helyes azonosítási arány (a szöveget teljesen német nyelvűnek feltételezve), majd a címkézés – szabály generálás folyamatát ismételve 70%-ra, majd 72%-ra emelkedett a helyes azonosítás.

A javulás mértékét a *6. ábrán* láthatjuk a három nyelvre. Az angol nyelvre a hibák egy jelentős része (10% illetve 14%) a nagyon hasonló skót nyelvre való tévesztésből fakadt. Figyelemre méltó, hogy annak ellenére növekedett ilyen mértékben a helyes azonosítási arány, hogy nem a szövegek tényleges nyelvére vonatkozó valószínűségeket használtunk, hanem az egymás mellett lévő szavak nyelvének egyezésére vonatkozó valószínűségeket. Az 1. táblázat alapján ez a szószintű azonosítási arány 80-90%-os helyes mondatszintű azonosítást tesz lehetővé a többségi döntés használatával.

Az Interneten elérhető különböző nyelvű szövegek nagy mennyisége miatt természetesen nem vagyunk rászorítva hogy ilyen kisméretű tanító halmazt használjunk, ezért gyakorlati alkalmazásokban a táblázat első sorában láthatónál is jobb, 100%-ot erősen közelítő helyes azonosítással számolhatunk.



6. ábra
A nyelvi valószínűségek figyelembevételével elért javulás

5. Összefoglalás

A cikkben rámutattunk az automatikus címkézési módszerek, mint például a nyelvazonosítás és szófaji címkézés, használatának jelentőségére. Illusztráltuk a probléma fontosságát a szövegfelolvasó rendszerek, ezen keresztül a távközlési alkalmazások számára.

Áttekintettünk néhány, a nyelvazonosításra használt módszert, majd ezek egyes gyengeségeinek kezelésére bemutattunk egy új, kétféle feltételes valószínűséget használó, ezek értékét döntési fa tanításával becslő eljárást. A módszer hatékonyságát demonstráltuk a nyelvazonosítás címkézés feladatán. Az eredmények igazolják a megközelítés életképességét. A módszer várhatóan jól használható más problémák kezelésére, például szófaji címkézés morfológiai elemző nélküli közelítő megoldására.

Irodalom

[1] Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P., "The Design, Implementation and Operation of a Hungarian E-mail Reader", International Journal of Speech Technology, Vol. 3, Numbers 3/4, December 2000, pp.217–236.

[2] G. Németh, Cs. Zainkó, G. Kiss, M. Fék, G. Olaszy, G. Gordos: "Language Processing for Name and Address Reading in Hungarian", Proc. of IEEE Natural Language Processing and Knowledge Engineering Workshop, Oct. 26-29, Beijing 2003, China, pp.238–243.

[3] Pfister, B., Romsdorfer, H., "Mixed-lingual text analysis for polyglot TTS synthesis", Proc. of Eurospeech 2003, pp.2037–2040.

[4] Halácsy, P., Kornai, A., Németh, L., Rung, L., Szakadát, I., Trón, V., "Creating open language resources for Hungarian", Proc. of LREC 2004, pp.203–210.

[5] Marcadet, J. C., Fischer, V., Waast-Richard, C., " A Transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis", Proc. of Eurospeech 2005, pp.2249–2252.

[6] Prószéky, G., "Humor: a Morphological System for Corpus Analysis. Language Resources for Language Technology.", Proc. of the First European TELRI Seminar, Tihany 1995, Hungary, pp.149–158.

[7] Németh, L., Halácsy, P., Kornai, A., Trón, V., "Nyílt forráskódú morfológiai elemző" In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged 2004, pp.163–171.

[8] Ted Dunning, "Statistical Identification of Languages", Computing Research Laboratory, New Mexico State University, 1994.

[9] G. Németh, Cs. Zainko: "Multilingual statistical text analysis, Zipf's law and Hungarian Speech Generation", Acta Linguistica Hungarica 2002, Vol. 49 (3-4), pp.385–405.

[10] Prager, J. M.: Linguini, "Language Identification for Multilingual Documents", Proc. of the 32nd Annual Hawaii International Conf. on System Sciences, 1999, Vol. 1, p.2035.

[11] Tian, J., Suontausta, J., "Scalable neural network based language identification from written text", Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 2003, Vol. 1, pp.48–51.

[12] Häkkinen, J., Tian, J., "N-gram and Decision Tree-based Language Identification for Written Words", Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio Trento, Italy, 2001.

[13] W. B. Canvar, J. M. Trenkle, "N-gram based Text Categorization", Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, 1994. pp.161–176.

[14] Sproat, R., Riley, M., "Compilation of weighted finitestate transducers from decision trees" In: Association for Computational Linguistics, 34th Annual Meeting, Santa Cruz, Canada, 1996. pp.215-222.

[15] Suendermann, D., Ney, H., "Synther - a New M-Gram POS Tagger", Proc. of the NLP-KE 2003, Int. Conf. on Natural Language Processing and Knowledge Engineering, Beijing, China, 2003.

[16] Halácsy P, Kornai A., Varga D., "Morfológiai egyértelműsítés maximum entrópia módszerrel", Magyar Számítógépes Nyelvészeti Konferencia, 2005.

Beszéddetekciós módszerek
vizsgálata és optimalizálása gépi beszédfelismerő rendszerekhez

TÜSKE ZOLTÁN, MIHAJLIK PÉTER, TOBLER ZOLTÁN, FEGYÓ TIBOR, TATAI PÉTER

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
mihajlik@tmit.bme.hu

Lektorált

Kulcsszavak: küszöbszint-alapú beszéddetekció, beszédfelismerés, spektrális entrópia, zajbecslés, VAD

A cikkben a küszöbszint-alapú beszéddetekcióhoz használható paramétereket vizsgáljuk. Először a beszéddetekció küszöbérték-érzékenységét analizáljuk egy kisebb teszthalmazon a különféle paraméterek mellett, majd az eredmények és gyakorlati megfontolások alapján választjuk ki a beszédfelismerési tesztekhez használt detekciós módszert. Az energia helyett a jóval robusztusabb spektrális entrópiát használjuk a beszéd jelenlétének kijelölésére. További különlegessége és újdonsága a megközelítésnek, hogy az entrópiaszámítás előtt spektrális részsáv-energiákon alapuló zajspektrum becslést használunk a zaj fehérítésére. Ennek eredményeképp nagymértékben zajtűrő, entrópia-alapú beszéddetekciós módszert kaphatunk. Ezen állításunkat számos beszédfelismerési kísérlettel támasztjuk alá, amelyekben normál, illetve kifejezetten zajos telefonbeszéd-felismerést végeztünk. A javasolt beszéddetekciós eljárás alkalmazásával minden esetben javult a felismerési pontosság (maximálisan 29,5%-kal), valamint a felismerendő keretek száma is jelentősen csökkent mind tiszta, mind zajos felvételek esetén.

1. Bevezetés

A beszéd-alapú szolgáltatások egyre növekvő száma szükségessé teszi hatékony, zajtűrő beszéddetektorok fejlesztését. A beszéd jelenlétének kijelölése igen fontos például a beszédfelismerőknél és a kissebességű beszédátvitel során.

Az előbbi esetben, hatékony beszéddetektálás esetén, a felismerő csak a beszédet tartalmazó kereteket kapja meg, így a felismerő beszédszünetekben kikapcsolható. A felismerés pontosabbá válhat, mert ilyenkor a nem-beszédet – amit általában a felismerő nem, vagy csak korlátozott mértékben tud kezelni – a rendszer nem próbálja a betanított szavak valamelyikéhez hasonlítani, ezáltal a felismerő hatásfoka javul, ráadásul a számításigény is csökken. Tehát egy jó beszéddetektor képes a beszédfelismerő rendszerek pontosságán és működési sebességén javítani.

A második esetben, a beszédátvitel során, a beszéddetektálás közismerten azért fontos, mert sávszélességet spórolhatunk meg, ha a csatornát beszédszünetekben nem foglaljuk. A távközlésben használt beszéddetektálási algoritmusok azonban közvetlenül nem használhatók a beszédfelismerésben, mert elsősorban nem a beszéd, hanem inkább a csend kijelölése a feladatuk, így nem szűrik hatékonyan a beszédfelismerést zavaró, nagyszintű zajokat.

Az elmúlt évek során számos detektálási algoritmust dolgoztak ki a beszédfelismerés számára. Ezek az eljárások többé-kevésbé két kategóriába sorolhatók [1]. Az első típusú algoritmus, úgynevezett **küszöbalapú** [1,2,9,11]. Ebben az esetben a bejövő jelből beszéd/nem-beszéd eldöntésére alkalmas paraméterek kinyerése után adaptív, az idővel változó, a környezethez alkalmazkodni próbáló vagy globális, előre beállított küszöbérték szerint történik a detektálás.

A küszöb-alapú beszéddetektálás legfontosabb lépései a következők:

– *Paraméter kinyerés:*
  olyan jellemzők előállítása a jelből, amelyek értéke más a zaj- és más a beszédszakaszokon.
– *Küszöbszint beállítás:*
  ennek alapján ítélhető meg egy jelszakaszról, hogy azt beszédnek vagy szünetnek tekintsük. Lehet adaptív vagy állandó is.

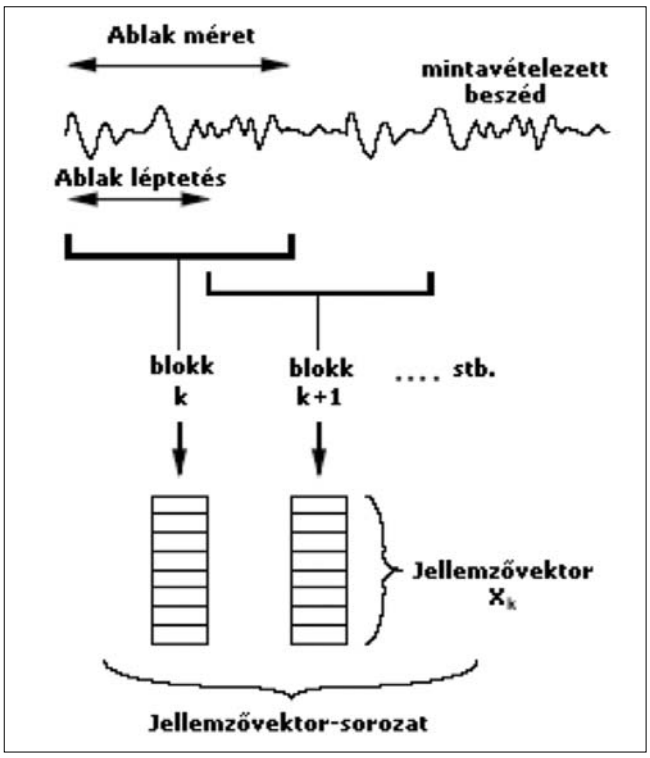
A másik elterjedt megközelítés a **mintaillesztésen alapuló beszéddetektálás** [4]. Ebben az esetben nemcsak a beszédről, hanem a zajról is modellt kell alkotni, és ennek paramétereit megbecsülni. A detektálás hasonlóan történik, mint maga a felismerési folyamat. A küszöb alapú módszert alkalmazó detektorokkal összehasonlítva, a mintaillesztésen alapuló eljárások tanító adatokat és nagyobb erőforrásokat igényelnek.

A továbbiakban a küszöb alapján döntő detektorokról lesz szó. Alapvetően egyszerűbbek és gyorsabbak, és jóval szélesebb az alkalmazási körük. Bár a dolgozatban elsősorban a beszédfelismerés hatásfokának javítását célozzuk a zajrezisztens beszéddetekcióval, a lehetséges alkalmazások túlmutatnak a beszédfelismerésen.

2. Beszéddetekciós paraméterek

A jelből olyan paramétereket célszerű kinyerni, amelyek különböző eloszlást mutatnak a beszédre és a nem-beszédre. Az egyes állapotok eloszlásának mérésére megfelelő adatbázis szükséges, az adott felvételeket pontosan fel kell címkézni.

A beérkező jel *k*. szakaszából *L* dimenziós paramétert kinyerve áll elő X_k paraméter-*oszlopvektor*. A detektálás során a paramétervektor alapján történik a



1. ábra Paramétervektor előállítása a jelből

döntés az előre felvett állapotok valamelyikére (H_i): a beszédre és a nem-beszédre. Ha az állapotok számát illetően csak kétféle osztályozás történhet (H_0, H_1), akkor a döntés a következő formában írható:

$$P(X_k|H_0)\cdot P(H_0) \underset{H_1}{\overset{H_0}{\gtrless}} P(X_k|H_1)\cdot P(H_1) \qquad (1)$$

ahol H_0: az aktuális keret nem-beszéd,
H_1: az aktuális keret beszéd.

Átrendezve és $\frac{P(H_0)}{P(H_1)}$ helyett más küszöböt, η-t választva, skálázhatóbbá válik a detektálás.

$$\frac{P(X_k|H_1)}{P(X_k|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (2)$$

Többdimenziós X_k esetén az *i.* állapothoz tartozó eloszlást általában Gauss-eloszlással közelítik. Egydimenziós paraméterek esetén könnyen mérhető és ábrázolható az (1)-es képletben szereplő, egyes állapotokra jellemző eloszlás sűrűségfüggvénye. A kétállapotú döntés miatt a küszöbérték kiindulási értékének a beszédhez és a nem-beszédhez tartozó paraméter-eloszlásfüggvények metszéspontja tekinthető. Ekkor az aktuális keretben mért paraméterérték alapján igen egyszerűen dönthetünk beszédre, illetve nem-beszédre.

2.1. Energia

Az energiaküszöb-alapú megközelítés előnye, hogy a zaj karakterisztikáját nem kell ismerni. Hátránya viszont, hogy érzékeny a nagy energiájú zajokra, hiszen nem minden beszéd, aminek nagy energiája van, azaz jelentősen csökkenhet a detekció hatékonysága. Alacsony jel-zaj viszony (SNR – Signal to Noise Ratio) esetén pedig a halk beszédszakaszok energiáját teljesen elfedheti a zaj energiája. Tehát az energia-alapú algoritmusok rossz eredményeket mutatnak zajos körülmények között. Az aktuális, T minta hosszú t_0. kezdetű keretben (ahol mintavett, azaz diszkrét idejű jelet dolgozunk fel) az energiát a következő módon számoljuk:

$$E_{jel}(t_0) = \sum_{t=t_0}^{t_0+T-1} y^2(t) \qquad (3)$$

A küszöbszint beállítása többféle módon lehetséges, például csúszó ablakos energiaátlagolással, vagy a t_0-t megelőző rövid időintervallumból a minimális energiaszintet választva. Beszédnek pedig azokat a szakaszokat tekinthetjük, amelyek energiája – például min. *6* dB-lel – a küszöb fölé emelkednek. A fentebb vázolt esetben nincs szükség spektrumszámolásra, aminek számottevő az erőforrás igénye. Bár létezik a spektrum alapján számolt energia-alapú detektálás is, a spektrumból más paraméterek is kinyerhetők és használhatók az energia mellett, illetve helyette.

2.2. Spektrális entrópia

E jellemző kiszámolásához szükség van a jel spektrumára. A beérkező jelet átlapolódó blokkokra bontva és e blokkokon FFT-t (Fast Fourier Transformation) végrehajtva kapjuk a jel gördülő spektrumát:

$$Y_{jel}(f,t_0) = \sum_{t=0}^{T-1} y(t_0+t)\cdot h(t)\cdot e^{-\frac{j2\pi\cdot t\cdot f}{T}}, \qquad (4)$$

ahol: t: a diszkrét idő,
$y(t)$: a vizsgált jel,
f: frekvencia,
t_0: az aktuális keret kezdete,
$h(t)$: a súlyozó ablak (általában Hanning).

Amíg a jel-zaj viszony elég nagy, addig az energia-alapú detektálás jól használható, de *SNR <0* dB esetén az eredmények már elég rosszak, noha a spektrumban még jól látszanak a beszédszakaszok, vagyis a spektrum még mutat bizonyos rendezettséget. A spektrum rendezettségének mérésére az információelméletből ismert Shannon-i entrópia mintájára [11] bevezeti az amplitúdó spektrum entrópiáját. Ezt az alábbiak szerint definiálja. Az információ-forrás entrópiája (Shannon) [9]:

$$H(S) = -\sum_{i=1}^{N} P(s_i)\cdot ld\{P(s_i)\}, \qquad (5)$$

ahol s_i a forrásból érkező *i.* szimbólum, $P(s_i)$ az *i.* szimbólum adási valószínűsége. Ezek alapján a *t.* keret *f* frekvencián kiszámolt spektrumának entrópiája [11]:

$$H\left(|Y_{jel}(f,t)|^2\right) = -\sum_{f=1}^{F} P\left(|Y_{jel}(f,t)|^2\right)\cdot ld\left\{P\left(|Y_{jel}(f,t)|^2\right)\right\}, \qquad (6)$$

ahol:

$$P\left(|Y_{jel}(f,t)|^2\right) = \frac{|Y_{jel}(f,t)|^2}{\sum_{f=1}^{F}|Y_{jel}(f,t)|^2}. \qquad (7)$$

Az entrópia egy véletlen változó bizonytalanságát írja le. Mivel a beszéd és a zaj más-más spektrális karakterisztikával rendelkezik, az entrópia alkalmas paraméterválasztásnak tűnik a beszéddetektálás döntési kritériumához.

Az entrópia maximális, ha a vizsgált jel fehérzaj, H_{max}= *ld(F)*; és minimális, ha a jel tiszta szinusz, H_{min}= *0*. Fontos, hogy az entrópia értéke a jelszinttől független. Így változó szintű, de állandó spektrális karakterisztikájú zaj esetén a beszédszakaszok az entrópiából könynyen kijelölhetők. A küszöb meghatározható adaptívan, de létezik statisztikusan becsült megoldás is [11].

Természetesen, ha növeljük a zajszintet, akkor a beszédkeretre számolt entrópia is változik, a zaj spektruma fokozatosan elnyomja a beszédét, a spektrum végül teljesen egyenletessé válik, és nem mutat rendezettséget *(2. ábra)*.

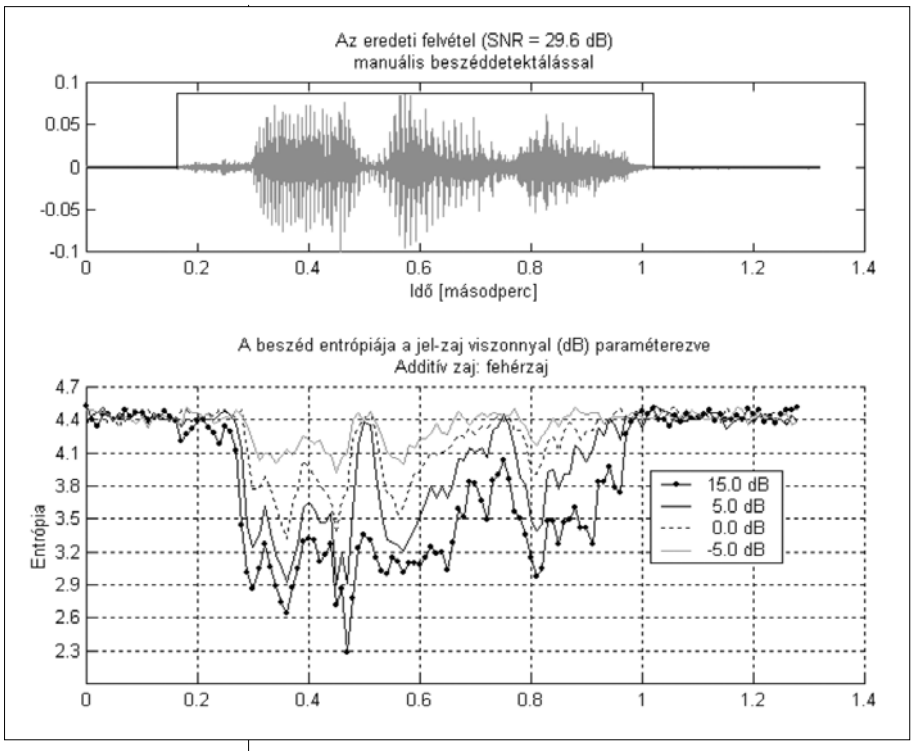
A spektrális entrópiaküszöbmódszer tehát jól használható beszéddetektáláshoz, ha a zaj fehér, azaz a spektruma egyenletes. Színes zaj esetén a zaj spektruma is rendezettebb, ezért nem lesz olyan egyértelmű a beszéd jelenléte az entrópia-idő diagramon.

A [11] irodalom az entrópia-alapú detekció egyéb zajokra való kiterjesztéséhez a következőt javasolja. Az aktuális keret spektrumát az entrópia számolása előtt osszuk el a *T* időre átlagolt spektrummal (8):

$$Y_{átlag}\left(f,t_0\right)=\frac{Y(f,t_0)}{\frac{1}{T}\sum_{t=-T/2}^{T/2}Y(f,t)}.$$

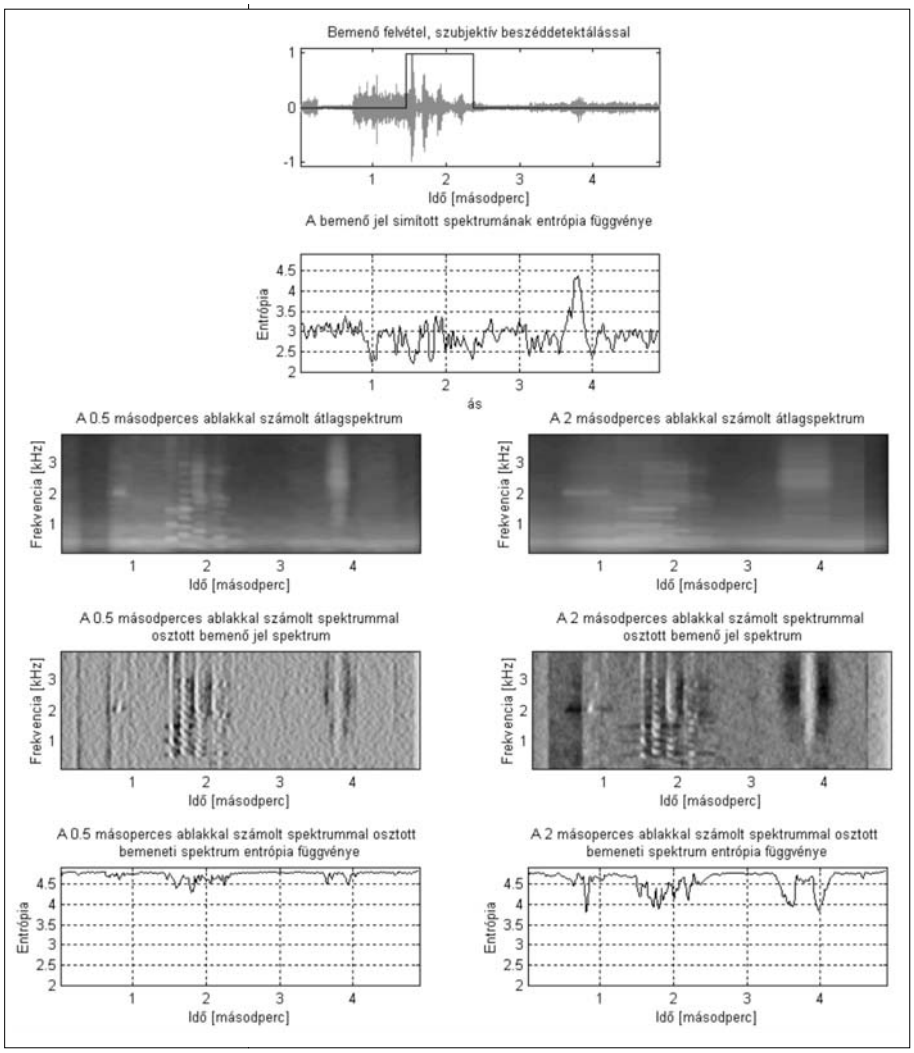
Az így kifehérített spektrumra számoljuk ki az entrópiát, és így a fehérzajnál alkalmazott detektálási módszer ebben az esetben is használhatóvá válik.

Tapasztalatunk szerint a beszédszakasz spektrumát a körülötte számolt átlagspektrummal osztva lerontjuk a beszéd entrópiáját is. Tehát a zaj spektruma



2. ábra Beszédjel entrópiájának alakulása fehérzajban

3. ábra Az entrópia alakulása átlagspektrummal való osztás hatására



valóban kifehéredik, de tulajdonképpen a beszéd spektruma is. Így a fehérzajnál alkalmazott detektálási módszer nem lesz elég eredményes színes zaj esetén *(3. ábra)*. A fenti eljárással az a probléma, hogy az átlagspektrum mindig tartalmazza a beszédspektrumot is, így az azzal való osztás mindig fehérítést jelent a beszédszakasz számára.

Természetesen adódik, hogy ha ismerjük a zaj – legalább közelítő – spektrumát, és a (8) nevezőjében az átlagspektrum helyett azt alkalmazzuk, akkor csak a zajspektrum fehéredik ki. Meglehet, hogy a beszédspektrum torzul ilyenkor, azonban a rendezettsége megmarad, így az entrópiája is alacsony marad, ugyanakkor a nem-beszéd szakaszok entrópiája közel maximális lesz. Ehhez tehát szükség van a beszéd alatti zaj spektrumának becslésére.

2.3. Hosszúidejű spektrális divergencia

A [8] alapján, ha ismerjük a jel gördülő amplitúdóspektrumát, $X_{k,l}$-t, ahol *k* a diszkrét időt, *l* a frekvenciasávot jelöli, akkor a jel *N*-ed rendű hosszúidejű spektrális „burkolója” (LTSE – Long-Term Spectral Envelope):

$$LTSE_N(k,l) = \max\left\{X_{k+j,l}\right\}_{j=-N}^{j=+N} \tag{9}$$

A *k*. keret hosszúidejű spektrális divergenciáját a (10) szerint kapjuk meg, az időben átlagolt zaj-amplitúdóspektrummal ($X_{Noise}(l)$) osztott *LTSE(k,l)* frekvenciakomponenseiből képzett átlagnak a logaritmusával (L jelöli a frekvenciasávok számát):

$$\mathrm{LTSD}_N(k) = 10\log_{10}\left(\frac{1}{L}\sum_{l=0}^{L-1}\frac{LTSE^2(k,l)}{X_{Noise}^2(l)}\right) \tag{10}$$

A képlet hasonló ahhoz, mintha minden frekvenciakomponensen jel-zaj viszonyt mérnénk, és átlagolnánk ezeket. A [8] állítása szerint ez a tényező egészen mást mutat zaj és mást beszéd esetén. A jó eredményhez persze szükség van a zaj spektrumának becslésére.

2.4. LPC együtthatók

A jelből kinyerhető LP (Linear Prediction) együtthatók alkalmasak a beszédspektrum burkolójának kinyerésére, a beszéd átvitele során lényegkiemelésre és tömörítésre. Az LPC együtthatók alapján történő beszédtömörítés alapja, hogy a beszéd spektrumát csak pólusokkal is jól lehet közelíteni, hiszen a zöngés hangok alap- és felharmónikus frekvenciái megfeleltethetőek az LPC-ből képzett szűrő pólusainak. Az *X(n)* jelből lineárisan predikált $X_p(n)$ jel alakja (11):

$$X_p(n) = -a_1 \cdot X(n-1) - a_2 \cdot X(n-1) - \ldots - a_N \cdot X(n-N)$$

Az a_k együtthatók meghatározása a becslés négyzetes hibájának, $\sum(X(n) - X_p(n))^2$, minimalizálásával történik. Az LP szűrő az „$X(n)$-$X_p(n)$”-t hibajelet állítja elő, és a *z* tartományban a következő módon írható:

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}}. \tag{12}$$

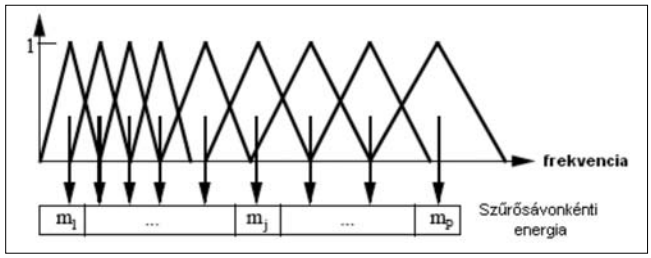
Az LP együtthatók alapján készült szűrő tehát jellemzi a beszéd spektrumát.

2.5. Mel-kepsztrum

Az amplitúdóspektrum egyenletes frekvenciaosztásokkal tartalmazza az adott keret energiájának eloszlását. Az emberi hallás azonban nem egyformán érzékeny az egyes frekvenciaközökre. Az emberi hallás frekvenciában nemlineáris karakterisztikáját figyelembe vehetjük, ha az adott keret spektrális energia-eloszlását lineáris Mel-skálán számoljuk ki. Az *f* frekvencia megfelelője a Mel-skálán [15]:

$$M(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right). \tag{13}$$

A nemlineáris karakterisztika megvalósításának egyik módszere az, ha a jelet időben szűrjük Mel-skála szerint elosztott sávszűrőkkel, és a sávokra különkülön számoljuk keretenként az energiát. A másik, és a gyakorlatban inkább használt módszer, ha az aktuális jelszakaszt Fourier-transzformáljuk, majd az egyes szűrősávokra eső energiát összegezzük a megfelelően változó számú frekvenciakomponensekre.



4. ábra Melszűrőbank és a szűrőnkénti energiák

A beszédfelismerésben azonban tipikusan nem a Mel-spektrumot, hanem annak egy származtatott menynyiségét, a Mel-kepsztrumot használjuk. Ebben az esetben az együtthatókat a jel Mel-skálás reprezentációjából DCT (Discrete Cosinus Transform) használatával nyerjük. Az *i*. MFCC (Mel-Frequency Cepstral Coefficient) együttható képlete:

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg(m_j)\cos\left(\frac{\pi \cdot i}{N}(j-0.5)\right). \tag{14}$$

Itt *N* a Mel-szűrőbankok száma, m_j a *j*. Mel-szűrőn mért energia az aktuális keretben. A Mel-kepsztrum együtthatókból általában nincs szükség az összesre, csak az első *p* darabra (általában p=12).

2.6. Kepsztrális divergencia

A [12] irodalom bevezeti a kepsztrális koefficienst *V*-t, ami nem más, mint a jel kepsztrális együtthatói négyzetének összege, azaz a keretenkénti kepsztrális együtthatók második momentuma.

$$V_1 = \frac{1}{D}\sum_{i=1}^{D} c_i^2. \tag{15}$$

3. Zajbecslés

A beszéd detektálásához mindig szükség van valamilyen beszédjellemző paraméterre, amelyekről az előző fejezetben adtunk áttekintést. Azonban a zaj-rezisztens

beszéddetekcióhoz általában szükség van még a zajjellemzők (tipikusan a zajspektrum) becslésére is.

[7] utal egy olyan fajta zajbecslésre, ami az időben visszatekintve minden frekvencia-komponensnek a minimumát ragadja ki. Az alapgondolat, hogy a beszéd gyorsan ingadozik, szünetekkel tagolt, így megfelelően nagy *T* időintervallumban a frekvenciakomponensek minimumát kigyűjtve csak a zajra jellemző spektrumot kapjuk, ha a zajt lassabban változónak tekintjük, mint a beszédet. A t_0 időponthoz tartozó becsült zaj spektrumát a következő módon kapjuk:

$$Y_{zaj}(f,t_0) = \min_{t=t_0-T\ldots t_0}\{Y_{jel}(f,t)\} \qquad (16)$$

Azonban könnyen belátható, hogy az újonnan belépő zajokkal szemben az eljárás tehetetlen, ezért az általunk javasolt zajbecslés nem csak a múltból, hanem a „jövőből" is vesz mintát a zajspektrum számításához. Természetesen a jövőbeni keretek spektrumának kiszámítása és felhasználása csak késleltetés árán történhet meg.

A becslés hatásosságának növelésére a becsléshez használt időintervallumot két részre bontottuk: T_1, illetve T_2 hosszú szakaszokra. Mindegyikben külön-külön történt a zajbecslés, azaz két zajbecslővel. Majd a két becsült zajspektrum frekvenciakomponensei közül mindig a nagyobbikat választva határoztuk meg az aktuális keretre vonatkozó zaj spektrumát. A becsült zaj t_0 időpillanatban tehát a következő (17):

$$\hat{Y}_{zaj}(f,t_0) = MAX\left[\min_{t=t_0-T_1\ldots t_0}\{Y_{jel}(f,t)\}, \min_{t=t_0\ldots t_0+T_2}\{Y_{jel}(f,t)\}\right]$$

T_1 és T_2 értékét akkorára érdemes választani, hogy a minimumot kereső ablakban bekövetkezzen beszédhangváltozás, vagyis az amplitúdóspektrum átrendeződése.

Például egy felpattanó zárhang előtt valószínűleg minden frekvenciakomponens minimumot fog elérni. A múltban működő zajbecsléshez hosszabb időintervallumot érdemesebb használni, mint a jövő mintáiból való zajbecsléshez, mert ez nem okozhat késleltetést. Viszont a jövőből hosszabb szakaszt venni csak akkor érdemes, ha az algoritmus adatbázison fut, mert valósidejű alkalmazásoknál megengedhetetlenül nagy késleltetést vihetünk be a rendszerbe, ha túl nagy az előretekintés.

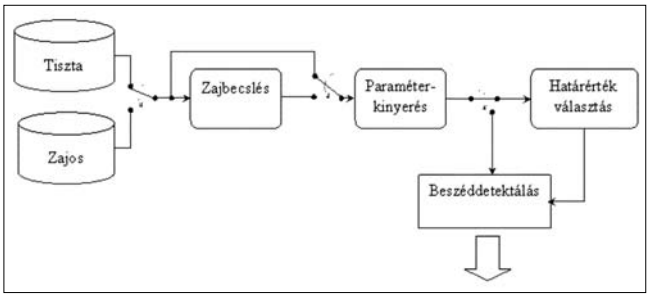
4. A beszéddetekciós paraméterek összehasonlítása ROC görbékkel

A szakirodalom a beszéddetektorokat az úgynevezett ROC (Receiver Operating Characteristics) görbékkel jellemzi, hasonlítja össze egymással. Ennek lényege, hogy a küszöbérték függvényében ábrázoljuk a detektálási eredményeket, a „mindent beszédnek detektálástól" a „mindent szünetnek detektálásig".

A grafikon x tengelyén a nem-detektált beszédszakaszok arányát (False Alarm Rate H0 = FAR0), az y tengelyen pedig a helyesen detektált szünet arányát (Hit Rate H0 = HR0) ábrázoljuk. Adott küszöb mellett ez meghatároz egy (x,y) pontot. A különféle küszöbszintekhez tartozó pontok összessége adja meg a vizsgált paramétert használó beszéddetektor ROC görbéjét. Ebben az értékelési módban nem játszik szerepet a beszéd- és a nem-beszédkeretek egymáshoz viszonyított mennyisége. Az a jobb detektor, amelyik a (0,1) ideális pontot minél jobban megközelíti, illetve amelyik ROC görbéje a nagyobb.

A 2. fejezetben említett beszéddetektálási paramétereket az *5. ábrán* látható mérési elrendezésben teszteltük.

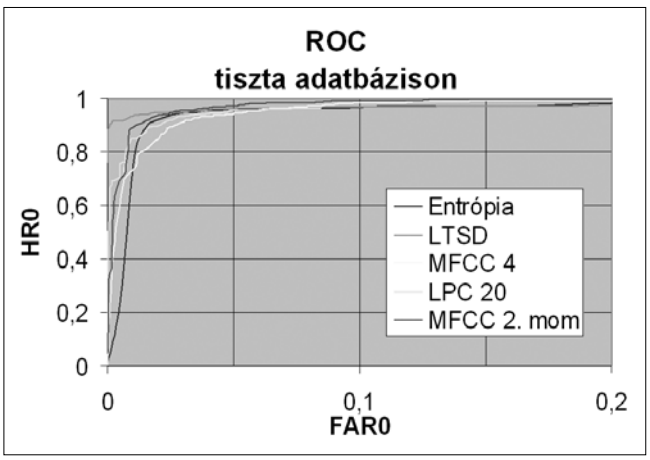
Az egyes paramétereket külön-külön optimalizáltuk. A zajos beszédanyag a [6] adatbázis 100 felvételből álló részhalmaza (mindegyik felvétel más beszélőtől származik). A tiszta adatbázis pedig az [5] adatbázis nem publikus tüköradatbázisának (Besztel) szintén 100 beszélőtől származó részhalmaza. Mindkét adatbázis valóságos (nem laboratóriumi) környezetben felvett telefonbeszédet tartalmaz, de az első esetben a beszélők kifejezetten arra lettek kérve, hogy zajos helyről telefonáljanak, míg a második esetben az utólag zajosnak minősített felvételeket nem válogattuk be a tesztalmazba.

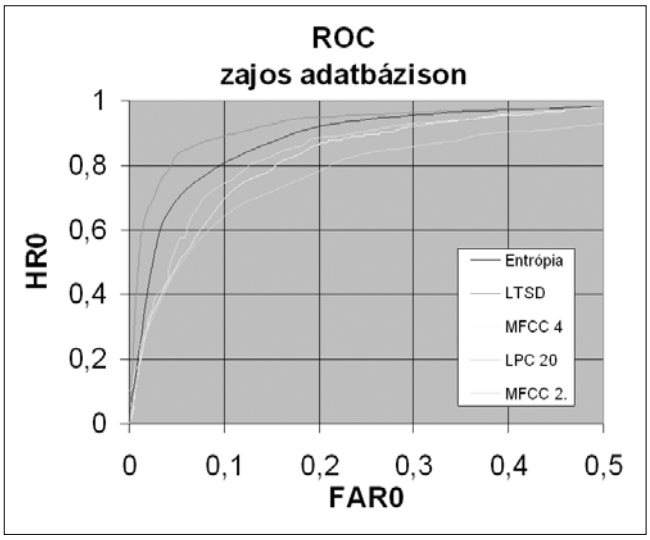


5. ábra A mérési elrendezés

A zajos és a tiszta adatbázisokon mért eredmények a *6. és a 7. ábrán* láthatóak. A mérések az entrópia, az LTSD, a 40 Mel szűrőből számolt 4 MFC együttható és az MFCC 2. momentuma típusú paraméterek esetén zajbecsléssel történtek. A 20 LP együttható pedig zajbecslés nélkül volt optimális.

6. ábra
Detektortípusok összehasonlítása tiszta adatbázison





6. ábra
Detektortípusok összehasonlítása zajos adatbázison

Mind a zajos mind a tiszta adatbázison mért ROC görbék világosan mutatják, hogy az LTSD paraméter a legmegfelelőbb a küszöbszint-alapú beszéddetekcióra a vizsgált paraméterek közül. Azonban az LTSD, vagyis a hosszú idejű spektrális divergencia számítása olyan nagy előretekintő időablakot igényel, ami az on-line rendszereknél nem engedhető meg. Így az előzetes ROC analízis második legjobban teljesítő jelöltjét, a zajbecsléssel korrigált spektrális entrópia-alapú beszéddetekciós megközelítést választottuk ki implementálásra és további beszédfelismerési vizsgálatokra.

5. A javasolt beszéddetekciós algoritmus

A bemutatandó beszéddetektor algoritmust NSSE-VAD-nak neveztük (Noise-Suppressed Spectral Entropy-based Voice Activity Detection, [14]), és a következő lépésekből áll (lásd a *8. ábrát*):

5.1. Gördülőspektrum-számítás

A bejövő jelet 30 ezredmásodperces keretekre bontva és Hanning ablakot használva, illetve 10 ezredmásodpercenként (a keretek 66,6% átlapolódásával) végzett Fourier-transzformálással számoltuk a spektrumot. Az összes beszédmintát fs = 8000 Hz-cel mintavételeztük.

5.2. Simítás

Frekvenciában simított spektrumon pontosabban végezhető a zajbecslés, jobban tükrözi a sztohasztikus jelek spektrumát. Például a fehérzaj spektruma ablakozás és Fourier-transzfromálás után nem konstans, míg simítás után jobban közelíti azt. A beszéddetektálást segíti, ha az entrópia görbe gyors időbeli ingadozásait kompenzálandó, időben simítjuk a gördülő spektrumot. A két művelet elvégezéséhez az amplitúdóspektrumot az idő és a frekvencia síkon egyszerre simítjuk.

Ehhez az alábbi S mátrix-szal adott kétdimenziós FIR szűrőt használjuk (18):

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \frac{1}{35}$$

$$Y_{simitott}(f_0, t_0) = \sum_{f=-2}^{2} \sum_{t=-2}^{2} Y_{jel}(f_0 + f, t_0 + t) \cdot S(f+3, t+3) \quad (19)$$

5.3. Zajbecslés

A zajbecslés a [7] által javasolt elgondolás továbbfejlesztett változata volt, ami (17) alapján úgy történt, hogy a zajbecslő késés nélkül képes volt követni a hirtelen belépő zajokat. A becsült zaj spektruma a minimum módszerből eredően nem lehet nagyobb egyik frekvencia-komponensen sem, mint az aktuális keret spektruma. A múltbeli zajbecslést a kísérleti tapasztalatok alapján T_2 = 0,75 másodpercre, a jövőbeli becslést pedig T_1 = 0,25 másodpercre választottuk.

5.4. Zajelnyomás

Az aktuális keret spektrumát (20) alapján fehérítjük. A jelspektrumból azért nem kivonjuk a zajt, mert ha az aktuális keret valódi zajának spektruma nem konstans, akkor a kivonás után a maradék spektrum sem fehér lenne, hiszen a becsült zaj csak kisebb lehet, mint a tényleges zaj. Ugyanakkor az aktuális keret spektruma a becsült zaj spektrumával való osztás után közel konstanssá válik. Tehát az entrópia a maximálishoz közeli lesz olyan keret esetén, amely beszédet nem, csak zajt tartalmaz.

$$Y_{zajelnyomott} = \frac{Y_{simitott}}{\hat{Y}_{zaj}} \quad (20)$$

5.5. Spektrális entrópia számítás

Az aktuális, becsült zajjal kifehérített keret spektrális rendezettségét $H(|Y_{zajelnyomott}(f,t)|^2)$-t a (6),(7) képletek segítségével számoljuk.

5.6. Elsőszintű döntés entrópiaküszöb alapján

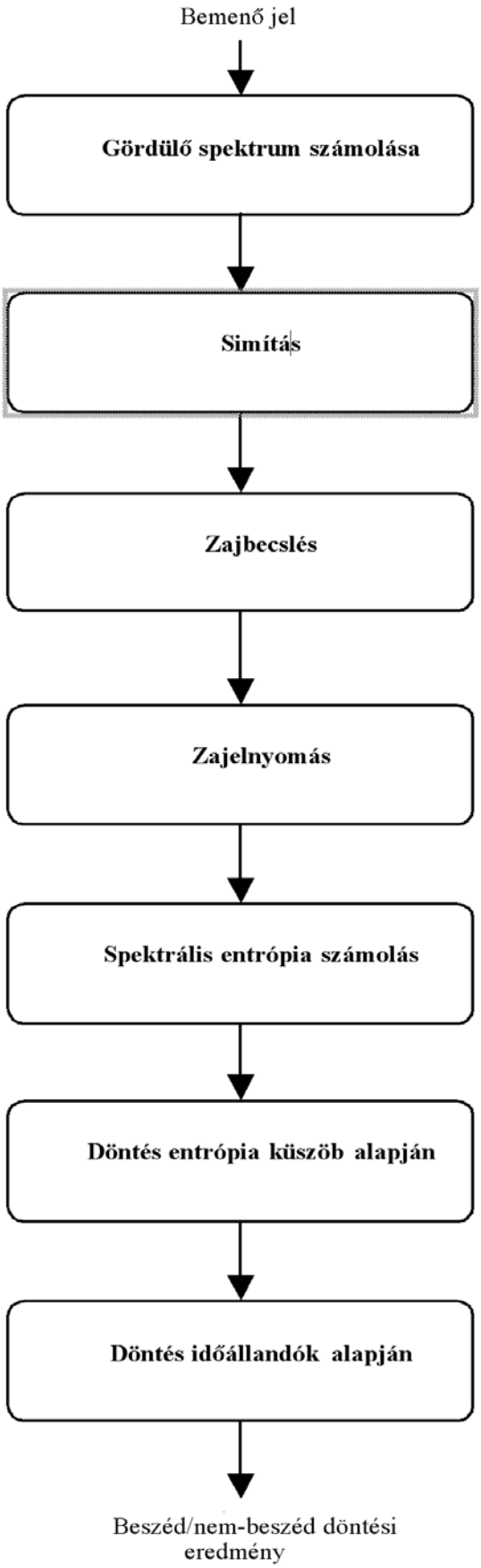
Az entrópia döntési küszöbét 4.5-nek választottuk. E felett zajnak, alatta beszédnek tekinti a detektor az aktuális keret. Fontos hangsúlyozni, hogy ez a fajta detektálási módszer globális küszöbön alapul. Nincs szükség adaptivitásra, ez a szerep a zajbecslőé. A küszöböt empirikus módszerekkel határoztuk meg.

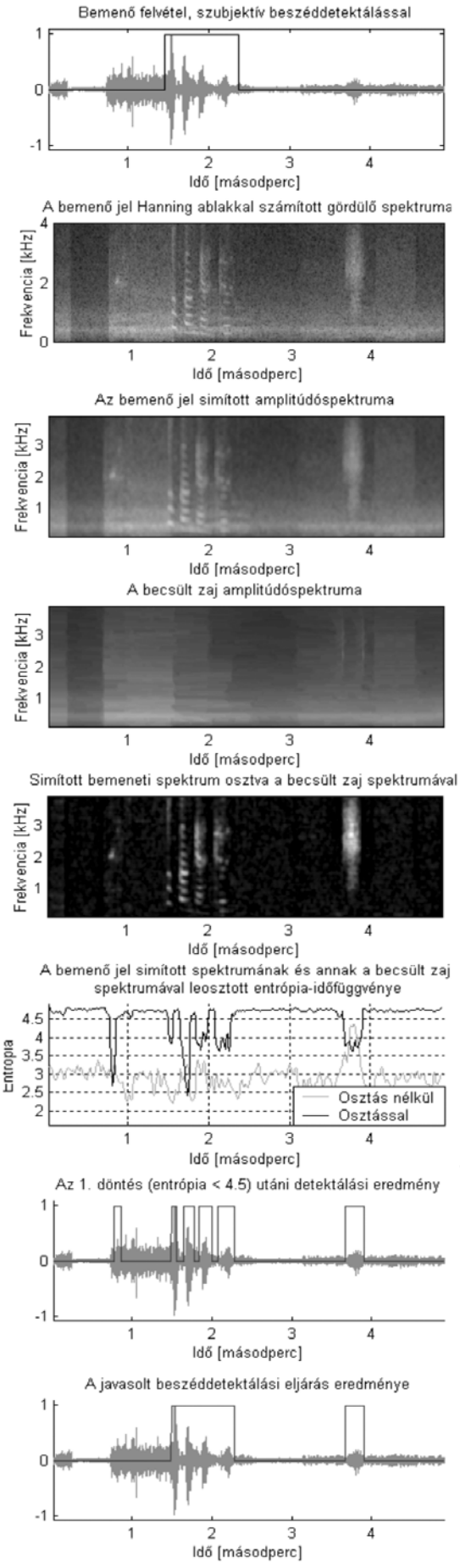
5.7. Második szintű döntés időállandók alapján

A beszédszakasz kijelöléséről az entrópiagörbe küszöb alá kerülésén kívül egy második réteg is dönt a következők szerint:

– A beszédszakasz minimális hossza
  0,2 másodperc,
  az ennél rövidebb beszédtartományok
  nem kerülnek detektálásra.
– A beszédben levő szünetek áthidalására
  a 0,1 másodpercnél kisebb időkülönbséggel
  rendelkező beszédszakaszok
  folyamatos szakaszként kerülnek kijelölésre.

8. ábra A javasolt detektor blokkvázlata és működése





6. Kiértékelés

Az ROC analízis, valamint a számos beszédfelvételen elvégzett szubjektív beszéddetekciós kísérletek eredményei jó okot adtak arra, hogy beszédfelismerő rendszerben alkalmazva is megvizsgáljuk a detektor működését, hatását a beszédfelismerésre.

A beszéddetekció hatékonyságát indirekt vizsgáltuk. A tanszéken alkalmazott, nyilvánosan is hozzáférhető beszédadatbázissal [5] betanított beszédfelismerő rendszer felismerési hibaarányát mértük különféle lényegkiemelő konfigurációs beállítások mellett.

6.1. Adatbázisok

Tanításra az MTBA (Magyar nyelvű TelefonBeszéd-Adatbázis) [5] kézzel szegmentált részét használtuk. A teszteléshez két másik telefonbeszéd-adatbázist vettünk igénybe. Elsőként az MTBA-hoz nagyban hasonló Besztel adatbázis **„tiszta”**, vagyis az annotáció során nem zajosként jelölt mintegy 6000 bemondását használtuk. A másik tesztadatbázisunk a nyilvánosan is hozzáférhető Tesztel [6], **„zajos”** telefonbeszéd adatbázis volt. Az ebben levő felvételek szándékosan természetes zajos környezetben (kocsiban, bevásárlóközpontban, utcán stb.), kifejezetten a zajtűrő beszédfelismerés vizsgálata végett készültek. Itt mintegy 1200 felvételt használtunk a tesztelésnél.

6.2. Vizsgálati módszer

Minden esetben 3 állapotú, „balról-jobbra” struktúrájú, környezetfüggő, rejtett Markov-modelleket használtunk hangmodellként. Mindkét tesztadatbázison parancsszó felismerést hajtottunk végre, a „tiszta” tesztadatbázison 1000 körüli szótárérettel, míg a „zajos” adatbázison 250 körüli szótárérettel, mindkét esetben a [13] felismerővel. Az azonos beállítású teszteket mindig párhuzamosan végeztük a két adatbázison. Tekintettel arra, hogy a zajos adatbázis felvételeinek jelentős része AGC (Automatic Gain Control)-torzított, minden beállításnál statikus energiával és anélkül is – az említett hatást kiküszöbölendő – elvégeztük a kísérleteket, így minden lényegkiemelési módszer esetén négy felismerési tesztet futtattunk. Végül nemcsak a javasolt detektort, hanem az ADSR (Advanced Ditributed Speech Recognition) ETSI szabványban rögzített detekciós eljárást is megvizsgáltuk.

6.3. Lényegkiemelési eljárások

A következő lényegkiemelési konfigurációk mellett végeztünk kísérleteket:

- Alkalmazva az ETSI ADSR lényegkiemelési szabványt, az abban foglalt jelalakformálást, zajelnyomást, vak csatornakiegyenlítést. (ADSR)
- Csak a Mel-frekvenciás kepsztrális együtthatókat számítva. (CC)
- A fenti mellett vak csatornakiegyenlítést is alkalmazva. (CC+BEQ)
- Csatornakiegyenlítést csak a teszteléskor végezve. (CC+fél BEQ)

6.4. Beszédfelismerési eredmények

Először beszéddetekció nélkül mértük az egyes konfigurációk hatásfokát.

Lényegkiemelő	Energiával		Energia nélkül	
	Tiszta	Zajos	Tiszta	Zajos
ADSR	**5,23**	**51,24**	**6,26**	**21,20**
CC	4,78	45,61	5,26	27,33
CC+BEQ	4,76	43,60	5,43	19,97
CC + fél BEQ	4,38	41,87	4,71	20,63

1. táblázat Referencia konfigurációk szó hibaaránya (WER – Word Error Rate, %) beszéddetektálás nélkül, zajos és tiszta adatbázison

Látható a referenciatáblázatban, hogy a statikus energia elhagyása igen jótékonyan hat a beszédfelismerés hatásfokára zajos esetben. Ez az AGC negatív hatásának kiküszöbölése miatt történhet. Ugyanakkor a tiszta adatokon kissé csökken a hatásfok.

Detektor	Lényeg-kiemelő	Energiával		Energia nélkül	
		Tiszta	Zajos	Tiszta	Zajos
ADSR	ADSR	5,21	51,07	6,26	21,20
NSSE	ADSR	5,11	36,14	5,86	20,54
NSSE	CC	4,66	35,51	5,08	22,77
NSSE	CC + BEQ	4,70	33,83	5,23	18,65
NSSE	CC + fél BEQ	4,27	30,94	4,51	18,48

Detektor	Lényeg-kiemelő	Energiával			Energia nélkül		
		Tiszta	Zajos	Átlag	Tiszta	Zajos	Átlag
ADSR	ADSR	+0,38	+0,33	**+0,36**	0,00	0,00	**0,00**
NSSE	ADSR	+2,29	+29,47	**+15,88**	+6,39	+3,11	**+4,75**
NSSE	CC	+2,51	+22,14	**+12,33**	+3,42	+16,68	**+10,05**
NSSE	CC + BEQ	+1,26	+22,41	**+11,83**	+3,68	+6,61	**+5,15**
NSSE	CC + fél BEQ	+2,51	+26,10	**+14,31**	+4,25	+10,42	**+7,33**

2. és 3. táblázat
A konfigurációk szó hibaaránya beszéddetektorokkal
A beszéddetektor által okozott relatív százalékos javulás

A következő mérési sorozatban pedig a javasolt NSSE-detektor által okozott hatást vizsgáltuk a beszédfelismerés szempontjából, valamint az eredményeket az ADSR saját beszéddetekciós eljárásának eredményeivel is összevetettük. Látható, hogy a javasolt detekciós algoritmus minden esetben javított a felismerési arányon. Különösen az energiát is tartalmazó zajos eredmények kimagaslóak (maximálisan 29,47%).

Bár a szóhiba-arány eredmények is ígéretesek az NSSE-VAD és az ADSR-VAD összehasonlítást illetően, a két beszéddetektor közti különbség drámaian megnő, ha a „nem-beszéd” keretek eldobási arányait tekintjük.

Adatbázis	Detektor	Vektorok száma	Keret dobási arány
Tiszta	ADSR VAD	1.788.101	24,9 %
	NSSE-VAD		60,0 %
Zajos	ADSR VAD	466.332	3,5 %
	NSSE-VAD		52,6 %

4. tábázat
A beszéddetektorok által a felismerés során az összes keretből eldobott keretek aránya %-ban

7. Összefoglalás

Többféle, zajtűrő beszéddetektáláshoz használatos paramétert vizsgáltunk meg. A ROC analízis alapján a praktikusan megvalósítható spektrális entrópia-küszöbön alapuló beszéddetekciós módszert választottuk ki implementálásra az általunk javasolt zajbecsléssel kiegészítve.

Megközelítésünket összevetettük az ETSI ADSR szabványában rögzített beszéddetekciós módszerrel. Az általunk használt, természetes háttérzajjal terhelt és háttérzaj-mentes telefonbeszédadatbázisokon a bemutatott detektálási algoritmus alkalmazásával egyrészt javultak a beszédfelismerési eredmények, másrészt az intenzív kereteldobás következtében jelentősen csökkent a felismerési folyamat erőforrásigénye. A zajbecslés az előretekintés miatt 0,25 másodperces késleltetést okoz, ami a valós idejű beszédalkalmazásoknál még megengedhető.

Irodalom

[1] Abdallah, I., Montrèsor, S., Baudry, M., „Speech signal detection in noisy environment using a local entropic criterion”, in Eurospeech, Rhodes, Greece, September 1997.

[2] Chuan JIA, Bo XU: Improved Entropy-Based Endpoint Detection Algorithm, ICSLP’02, Beijing, 2002.

[3] ETSI standard doc., ETSI ES 202 050 v1.1.1.

[4] E. Kosmides , E. Dermatas, G. Kokkinakis, „Stochastic endpoint detection in noisy speech”, SPECOM Workshop 1997., pp.109–114.

[5] http://alpha.ttt.bme.hu/speech/hdbMTBA.php

[6] http://alpha.ttt.bme.hu/speech/hdbtesztelen.php

[7] Izhak Shafran, Richar Rose: Robust Speech Detection And Segmentation For Real-Time ASR Application, Proc. of IEEE Int’l Conf. on Acoustic Signal and Speech Processing (ICASSP), Hong Kong, 2003. Vol.1, pp.432–445.

[8] Javier Ramírez, José C. Segura, Carmen Benítez, Ángel de la Torre, Antonio Rubio, „Efficient voice activity detection algorithms using long-term Speech information”, Speech Communication 42 (2004), pp.271–287.

[9] Jialin Shen, Jeihweih Hung, Linshan Lee, „Robust entropy based endpoint detection for speech recognition in noisy environments”, International Conf. on Spoken Language Processing, Sydney, 1998.

[10] Péter Mihajlik, Zoltán Tobler, Zoltán Tüske, Géza Gordos; Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, Eurospeech 2005, Lisbon.

[11] Philippe Renevey, Andrej Drygajlo: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech 2001, Aalborgh.

[12] Sergei Skorik, Frédéric Berthommier, „On a cepstrum-based speech detector robust to white noise”, SPECOM Workshop, St. Petersburg, 2000.

[13] T. Fegyó et al. „Voxenter – Intelligent Voice Enabled Call Center for Hungarian”, EUROSPEECH 2003. pp.1905–1908.

[14] Zoltán Tüske, Péter Mihajlik, Zoltán Tobler, Tibor Fegyó; Robust Voice Activity Detection Based on the Entropy of Noisesuppressed Spectrum, Eurospeech 2005, Lisbon.

[15] Steve Young at al.: The HTK Book, Cambridge, 2001.

Speech markup languages
Keywords: speech application, SALT, SRGS, SSML, VoiceXML

The last eight years have witnessed a shift from using proprietary approaches for developing speech enabled applications to using strategies and architectures based on industry standards. Developing speech and multimodal applications is assisted by a dozen XML-based markup languages. This article aims to survey speech standards for speech interaction, speech input, speech output and communication among speech components.

Experimental medicine information system with speech modules
Keywords: Profivox-Med medicine speech synthesizer, Latin exception-vocabulary, text understanding and extraction

The "Medicine line" is an automatic information system. Its aim is to enable Hungarian citizens to reach the necessary basic information about medicaments available in Hungary. There is no similar information system at present in Hungary. The National Institute of Pharmacy coordinates the approval of new drugs and issues the basic written information sheet about them in the form of package leaflets. This textual information will be provided by the system automatically. The number of different medicaments used in Hungary is about 5000. The system allows three level automatic, 24 hour information provisioning. The levels are: telephone (equipped with speech synthesis and recognition technology), Web and WAP,.

Hungarian continuous, medium-vocabulary speech recognizer: a medical transcription application
Keywords: HMM-models, n-gram models, perplexity

A development tool for constructing continuous speech recognizers (MKBF 1.0) has been created under Windows XP. The system is based on a statistical approach (HMM phoneme models, and bigram linguistic models with non linear smoothing). and works in real time. It can construct a medium-sized speech recognizer with vocabulary of 1000-20000 words. New solutions have been developed in the acoustical preprocessing, in the statistical model building of phonemes, and in syntactic level. Through our examination different training sets were used with different vocabularies. Hungarian is a strongly agglutinative language, where the number of the word forms is very high. This is the reason why two forms of the bigram linguistic model were constructed: one is the traditional word forms based and the other is the morpheme based model, where the vocabulary is much smaller. In this article, test results and the experiences drawn from them are presented. More than 91% recognition accuracy has been reached using perplexity based linguistic adaptation.

Change of generations in speech synthesis
Keywords: formant synthesis, unit concatenation, corpus based speech synthesis, speech quality evaluation

This paper gives an overview of the advancement of text-to-speech technologies over three generations. An experimental version of the corpus based unit selection text-to-speech system developed at BME TMIT is introduced.. The details of the automatic labelling of sound and word boundaries in the speech corpus are described. Different possibilities of prosody generation in a corpus based, unit selection synthesizer are explored. The procedure of unit selection in the system under development is described. The speech quality of the experimental system is compared to that of the earlier Hungarian text-to-speech systems.

Conversion of speech to facial animation to aid the communication of deaf people
Keywords: audiovisual speech processing, facial animation, multimodal communication, lip reading

A speech to facial animation direct conversion system was developed as a communication aid for deaf people. Utilising the preliminary test results a specific database was constructed from audio and visual records of professional lip-speakers. The standardized MPEG-4 system was used to animate the speaking face model. The trained neural net is able to calculate the principal component weights of feature points from the speech frames. The control coordinates have been calculated from PC weights. The whole system can be implemented in standard mobile phones. Deaf persons were able correctly recognize about 50% of words from limited sets in the final test based facial animation model.

Text corpus design for an automatic speech prompt generator
Keywords: interactive voice response systems, corpus-based speech synthesis, speech databases

The primary output of interactive voice response systems are the prompts – the systems use these prerecorded messages to inform the users of menu options, to acknowledge user actions, etc. The low entropy of prompt texts suggests that by using a dedicated speech synthesizer (a prompt generator), it is possible to synthesize messages with a naturalness near to human speech. In order to achieve this, the speech corpus of the prompt generator needs to be representative of the expected inputs so that the prompts can be concatenated from the least possible units. In this paper, we investigate this issue: after giving an overview of the system under development, we propose a method to design the text corpus of the prompt generator. Finally, we present a method to examine the representativity of the corpus by an independent collection of prompt texts.

Linguistic and phonetics aspects of corpus-based speech synthesis
Keywords: sound structures, optimal join points, invariance in prosody

Human speech is a unique and single product in time, containing also the special characteristics of the given language. A speaker cannot repeatedly produce exactly the same waveform even if she/he pronounces the same sentence. The actual state of her/his biological speech production system defines the produced speech wave. When synthesizing speech, always the same building elements (stored in a database) are used. The contradiction between these two facts cannot be solved fully by present technologies but the the percieved effect may be reduced. In this article we summarize the main linguistic and phonetic aspects of speech that can help in achieving this goal.

A machine learning algorithm for text labelling and its' application for speech synthesis
Keywords: machine learning, language identification, LID, TTS

In this paper we present a novel machine learning approach usable for text labelling problems. We illustrate the importance of the problem for text-to-speech systems and through that for telecommunication applications. We introduce the proposed method, and demonstrate its effectiveness on the problem of language identification, using three different training sets and large test corpora.

Evaluation and optimization of speech end-point detection methods in automatic speech recognition systems
Keywords: Voice Activity Detection, spectral entropy, noise estimation and suppression

This paper deals particularly with the speech parameters applicable for threshold-based voice activity detection. First the detection accuracy vs. threshold value correspondences obtained using various speech features are analyzed. Then, as a compromise between practice and theory, the short time spectral entropy of the speech signal is chosen among the features. The novelty of our approach is the application of a pre-processing step before the spectral entropy calculation, where noise spectrum estimation and noise whitening is performed. As a result enormous improvement in speech end-pointing is observed in terms of accuracy and noise-robustness. Furthermore, the end-pointing method is tested in various speech recognition experiments, where normal and naturally noisy telephony speeches are used for the tests, as well. The proposed end-point detection approach achieves always positive, maximum rel. 29.5% word error rate reduction as compared to the "no-speech detection" case. In addition, the recognition process is significantly accelerated because about half of the speech frames are discarded thanks to the end-point detector.

Summaries • of the papers published in this issue

Contents

Cover: Reference points for visual speech synthesis on natural face and artificial model

Szerkesztőség
HTE Budapest V., Kossuth L. tér 6-8.
Tel.: 353-1027, Fax: 353-0451, e-mail: info@hte.hu

Hirdetési árak
1/1 (205x290 mm) 4C 120.000 Ft + áfa
Borító 3 (205x290mm) 4 C 180.000 Ft + áfa
Borító 4 (205x290mm) 4 C 240.000 Ft + áfa

Cikkek eljuttathatók az alábbi címre is
Szabó A. Csaba, BME Híradástechnikai Tanszék
Tel.: 463-3261, Fax: 463-3263
e-mail: szabo@hit.bme.hu

Előfizetés
HTE Budapest V., Kossuth L. tér 6-8.
Tel.: 353-1027, Fax: 353-0451
e-mail: info@hte.hu

2006-os előfizetési díjak
Közületi előfizetők részére: bruttó 30.450 Ft/év
Hazai egyéni előfizetők részére: bruttó 6.800 Ft/év
HTE egyén tagok részére: bruttó 3.400 Ft/év

Subscription rates for foreign subscribers:
12 issues 150 USD,
single copies 15 USD

www.hte.hu
Felelős kiadó: NAGY PÉTER
Lapmenedzser: Dankó András

HU ISSN 0018-2028

Layout: MATT DTP Bt. • Printed by: Regiszter Kft.